TailNLG: A Multilingual Benchmark Addressing Verbalization of Long-Tail Entities

Anonymous ACL submission

Abstract

Automatic verbalization of structured knowledge is a key task for making knowledge graphs accessible to non-expert users and supporting retrieval-augmented generation systems. Although recent advances in Data-to-Text generation have improved multilingual coverage, little attention has been paid to potential biases in the verbalization of rare entities, frequently known as long-tail entities. In this work, we present the first systematic study of long-tail entities in Data-to-Text generation. We introduce TailNLG[1], a new multilingual benchmark in English, Italian, and Spanish, built from Wikidata and covering entities with varying levels of popularity. We evaluated three different families of large language models in zero-shot settings and compared their performance on rare *versus* common entities, as well as against the established WebNLG benchmark. Our results suggest a tendency toward lower performance on long-tail entities: embedding-based scores are directionally lower for rare entities across most models, and within-model uncertainty as measured by Perplexity is consistently higher.

1 Introduction

Automatic conversion of structured information into human-readable text is a crucial task to make knowledge accessible to a broad audience. The information contained in knowledge graphs (KGs), catalogues, and taxonomies is typically accessible only to domain experts, whereas the general public requires it to be expressed in natural language. This need also extends to retrieval-augmented generation (RAG) technologies, which leverage textual knowledge bases to improve the safety and accountability of large language models (LLMs).

Despite recent advancements in this field and the growing number of multilingual resources (Ferreira et al., 2018; Shimorina et al., 2019; Cripwell et al., 2023; Oliverio et al., 2025; Ramón-Ferrer et al., 2026), there is a lack of work aimed at uncovering biases in the verbalization of structured knowledge (Blodgett et al., 2020). Specifically, an important open challenge is to better understand how LLMs handle rare entities, commonly referred to as as *long-tail* entities, and whether their performance differs from that on well-known entities. Although it has been demonstrated that LLMs exhibit performance drops when handling rare entities in tasks such as entity linking (Boscariol et al., 2024), representation of factual knowledge (Sun et al., 2024), and question answering (Hogan et al., 2025), there is no prior work that provides in-depth investigation of how different models handle long-tail entities in Data-to-Text generation.

In this work, we tackle this gap by presenting the first systematic study of long-tail entities in Data-to-Text verbalization (i.e., the transformation of structured data into natural language that faithfully conveys the semantics of the data in a human-

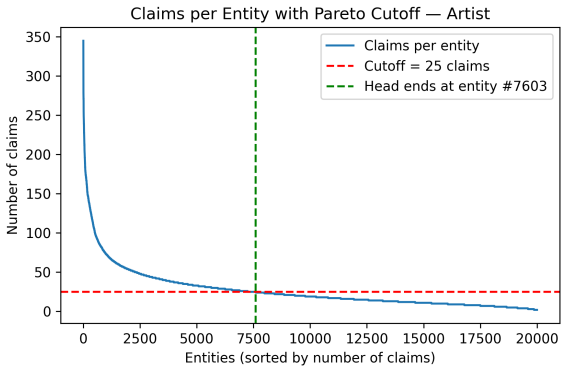

Figure 1: Distribution of entities by claims in the Wikidata category *Artist*. A small number of entities are associated with many claims, while the majority (long-tail) are associated with few relations. The red dashed line indicates the Pareto cut-off, while the green dashed line marks the head–tail threshold

[1]Full code and Benchmark available here https://anonymous.4open.science/r/TailNLG-benchmark-B339/README.md

understandable form). We focus on RDF-to-Text verbalization, as the Resource Description Framework (RDF) is a standard data model for representing information in Knowledge Graphs, encoding data as triples in the form ⟨subject, relation, object⟩. Recognising the socio-cultural factors that can determine the rarity of an entity (Adams et al., 2019; Stranisci et al., 2023), we adopt a multilingual approach. Our study introduces **TailNLG**, a novel multilingual RDF-to-Text benchmark covering three languages (English, Italian, and Spanish). The dataset contains triples extracted from Wikidata (Vrandečić and Krötzsch, 2014), including entities from different categories and spanning varying levels of popularity, from top-head (most well-known) to long-tail (less well-known) entities.

Through the creation of TailNLG, our work addresses the following research questions:

- **[RQ1]** Do LLMs perform differently in the verbalization of rare entities, and does language impact the verbalization output?
- **[RQ2]** Is the model uncertainty higher when dealing with rare *versus* common entities in the verbalization task, and how does model performance differ when comparing TailNLG with a well-established resource such as WebNLG?

To this end, three families of LLMs are assessed on the verbalization of long-tail entities in two settings: (i) comparing them with head entities within the TailNLG corpus, and (ii) comparing them with WebNLG (Gardent et al., 2017b), an established benchmark for the RDF-to-Text task. We assess performance using a diverse set of automatic measures, capturing semantic similarity, surface-level overlap with reference texts, and Perplexity (PPL) (Jelinek et al., 1977) to quantify model uncertainty. Our results suggest that LLMs have a systematic bias against long-tail entities: embedding-based metrics tend to be lower for long-tail entities and models' uncertainty in text generation of triple containing long-tail entities is consistently higher. In addition, language-specific bias also appear to influence the verbalization of long-tail entities. Analysis of PPL scores suggests that embedding-based and overlap-based metrics fail to systematically quantify their impact on verbalization, indicating that a more reliable evaluation framework of fairness in RDF-to-Text is needed.

2 Related Work

Long-tail entities are generally understood as those that occur with low frequency across large scale data sources, including training corpora (Kandpal et al., 2023), Wikipedia (Mallen et al., 2023), and large knowledge bases such as Wikidata (Kumar et al., 2024). Recent work has shown that LLMs exhibit a marked decline in performance when processing long-tail inputs (Graciotti et al., 2025; Hogan et al., 2025; Li et al., 2024; Sun et al., 2024) and that LLMs memorization is highly influenced by the frequency of information in the pretraining data (Kandpal et al., 2023). Several task-specific evaluations further confirm this pattern. Boscariol et al. (2024) find that entity linking is particularly difficult for rare entities when comparing various LLMs with ReLiK (Orlando et al., 2024), a state-of-the-art (SoTA) entity linking and relation extraction system. GRADES, an evaluation framework for graph-based question answering (Draetta et al., 2025), shows that SoTA models struggle with rare entities at multiple stages; in particular, verbalization suffers from limited semantic understanding of long-tail entities. In knowledge extraction, Graciotti et al. (2025) introduce KEMISTO, a benchmark centered on low-popularity entities, and demonstrate that LLMs perform significantly worse on it and exhibit systematic biased failures. A similar pattern emerges in question answering: using a benchmark spanning head, torso, and tail entities, Sun et al. (2024) show that accuracy consistently declines as entity frequency decreases, confirming that limited training exposure hampers LLMs' knowledge of long-tail entities.

Despite increasing attention to long-tail evaluation, no existing work specifically examines how well LLMs verbalize information when required to handle rare entities.

In the context of Data-to-Text, the task of verbalizing RDF triples in NLP has evolved substantially, driven both by template-based NLG systems and more recent neural approaches. Early work focused on hand-crafted or automatically induced templates, which provided high precision and strong control over the output, but suffered from limited scalability and reduced the ability to generalize beyond predefined patterns (Duma and Klein, 2013). With the advent of neural encoder–decoder architectures, particularly LSTM-based triple encoders and later transformer models,

RDF-to-Text generation became considerably more flexible, enabling fluent and information-rich verbalizations that better capture lexical and syntactic variation (Distiawan et al., 2018; Oliverio et al., 2024). Nevertheless, challenges related to factual consistency and robustness to unseen entities have motivated a series of enhancements to transformer-based systems. For example, integrating reward signals derived from information-extraction models has been shown to improve factual accuracy (Gao et al., 2021), while leveraging large external corpora, pre-training noise processes, and data augmentation strategies leads to more resilient generalization, particularly in low-resource or zero-shot scenarios (Montella et al., 2020; Zhang et al., 2023).

To support systematic development and evaluation of this task, the WebNLG benchmark has emerged as one of the primary datasets for RDF-to-Text, providing multiple versions, multilingual extensions, and low-resource configurations that have enabled extensive comparative studies across approaches and languages (Gardent et al., 2017b; Castro Ferreira et al., 2020).

3 The TailNLG Benchmark

Since no benchmark currently targets long-tail triple–verbalization pairs, we introduce TAILNLG, the first dataset specifically designed for long-tail verbalization. Inspired by Gardent et al. (2017b), we created the TailNLG benchmark following the same procedure employed by the authors, from the entity extraction strategy to verbalization, with the aim of facilitating the use of data by adhering to a standard structure. Constructing the benchmark in three languages adds another layer of complexity, as rare entities may not be consistently represented across multilingual resources. We prioritized high data quality by incorporating human evaluation and correction at each stage of the pipeline.

3.1 Entities Selection

To be consistent with previous resources, we followed the extraction methodology proposed by Gardent et al. (2017a) by selecting the same entity categories. Although there is not yet a widely accepted and unambiguous definition of long-tail entities (Jiang and Joshi, 2024; Boscariol et al., 2024; Kumar et al., 2024), following Sun et al. (2024) we took into consideration two main criteria in defining long-tail entities: (i) the number of claims (Wikidata relations) linked to a certain entity and (ii) the number of Wikipedia pages an entity has. Firstly, we set a threshold based on the number of claims linked to each entity. As noted by Kumar et al. (2024), large-scale resources such as Wikidata typically follow a Zipf-like distribution, where a small number of entities are associated with many claims while the vast majority with few. Based on a subset of $20,000$ entities for each category (complete list of categories in Table 4 in Appendix D) we define the long-tail threshold using a Pareto cut-off: we compute the minimum number of claims required for an entity to fall within the top 20% of entities that account for 80% of all claims (the head). Entities below this threshold are classified as long-tail. Since claim distributions differ across categories, we compute the cut-off separately for each category. Figure 1 shows an example of the threshold for the categories Artists: a small number of entities exhibit a very high number of claims, while the vast majority fall into a long-tail distribution. To further refine the set of long-tail entities below the Pareto cut-off, we additionally select those that have a Wikipedia page only in one of the three target languages (e.g., entities with Wikipedia page in English but not in Italian and Spanish).

In summary, we categorize entities into: (i) **long-tail entities**, whose number of claims falls below the Pareto cut-off, such as *Doğan Güzel* (Q268285), a Turkish cartoonist; and (ii) **Head entities**, whose number of claims exceeds the Pareto cut-off, such as *London* (Q84) or *Google* (Q95).

To examine model confidence on long-tail *vs.* head entities, assessing the reliability of the TailNLG dataset, we compute the average Perplexity (PPL) as a proxy for model uncertainty across LLMs of different families and sizes (see Section 4). For each model, the entity string is tokenized using the model's corresponding tokenizer, and PPL is computed as the exponential of the average negative log-likelihood of the resulting token sequence. Results in Figure 3 (Appendix A) show that long-tail entities exhibit higher PPL than head entities, indicating greater model uncertainty when processing long-tail entities. This trend supports the reliability of our entity extraction methodology and the overall quality of the dataset.

3.2 Triples Extraction

The next step was to extract data units for each entity. We define a data unit as a set of RDF triples, each consisting of a subject, a predicate, and an

object (e.g., Jet Li, place of birth, Beijing). To obtain diverse data units that vary in relation type and size, we followed Perez-Beltrachini et al. (2016) and extracted units with different relational configurations: Chain where the object of one triple serves as the subject of another, Sibling where the triples share the same subject and mixed where the data unit contains both sibling and chain relations. Some examples in appendix B.

For each QID (Wikidata entity ID), we automatically extracted, via a SPARQL query from Wikidata, a set of 10 data units per entity in each of the three languages. Since not all triples were available across all languages, we further filtered the data units by selecting only those for which triples existed in every language, thereby ensuring cross-lingual alignment and resulting in a set of parallel data units, as exemplified below:

> **EN:** `[Solar, creator, Paul S. Newman], [Paul S. Newman, place of birth, Manhattan], [Paul S. Newman, award received, Inkpot Award]`
> **ES:** `[Solar, creator, Paul S. Newman], [Paul S. Newman, lugar de nacimiento, Manhattan], [Paul S. Newman, premio recibido, Premio Inkpot]`
> **IT:** `[Solar, creatore, Paul S. Newman], [Paul S. Newman, luogo di nascita, Manhattan], [Paul S. Newman, premio ricevuto, Premio Inkpot]`

We obtained 1878 data units, divided into 1224 containing long-tail entities and 654 head. The complete distribution per category and type is in Appendix D.

3.3 Verbalization

To build the benchmark, we aimed to produce *<Data_Unit, Verbalization>* pairs for all data in all languages. Because manual verbalization is time consuming we adopted a two-step approach: we first manually verbalized a subset of data units for each language, then cross-translated them into the remaining languages using SoTA machine translation models in a human-in-the-loop setup.

The verbalization process was carried out on a subset of 689 data units by nine volunteer native speakers in the three languages. Annotators were proficient in at least two languages, allowing for a cross-check validation. They were instructed to verbalize the triples as naturally as possible, producing grammatically correct and fluent sentences while following guidelines designed to promote naturalness and ensure data homogeneity. All guidelines and some examples of verbalization are reported in Appendix C.1. Annotators could skip a data unit if they judged it incorrect or incomplete. After verbalization, a validation phase followed in which the annotators cross-checked each other's output, confirming or revising the initial sentences (full guidelines in Appendix C.2). Ambiguous cases were discussed collectively before a final decision was made.

3.4 Translation

The manual verbalization phase produced a total of 616 gold verbalizations (130 in English, 142 in Spanish, and 344 in Italian[2]). To create a fully parallel dataset we adopted an approach based on Machine Translation (MT) and Automatic Post-Editing (APE), as proposed by Aditya Hari et al. (2023). Drawing on earlier studies on cross-lingual extensions of triples-to-Text datasets (Oliverio et al., 2025; Ramón-Ferrer et al., 2026), we leveraged DeepL for Automatic Translation[3]. For each manually created verbalization in a source language (X) corresponding translations were automatically generated in the other two target languages (Y and Z), resulting in a total of 1,232 automatically generated translations.

To mitigate known issues in MT, such as omissions and hallucinations (Jurafsky and Martin, 2025), we introduced an APE phase. To identify errors in the translations, we leveraged the 3.5B version of xCOMET Guerreiro et al. (2024) [4], a multilingual model for quality estimation and error span extraction. For each input, the model outputs a sentence-level score (ranging from *Excellent* to *Weak*), as well as an detected error spans field, which identifies potential errors and their positions within the sentence. Subsequently, for the error correction phase, we adopted xTOWER[5], a multilingual decoder-only model proposed by Treviso et al. (2024), designed to provide natural-language explanations of errors present in a translation and suggest a corrected version of the sentence (example of prompt and output in Appendix F.1). We perform error detection and correction on all the MT translated data; the resulting output is a parallel dataset consisting of data units and verbalizations in Italian, Spanish and English as follows:

[2] This linguistic imbalance is due to the varying availability of expert native-speaking annotators.

[3] https://www.deepl.com/translator

[4] https://huggingface.com/Unbabel/XCOMET-XL

[5] https://huggingface.com/sardinelab/xTower13B

Source → Target	Accuracy	Fluency
$es \rightarrow en$	4.45 ± 0.58	4.99 ± 0.07
$en \rightarrow es$	4.57 ± 0.53	4.95 ± 0.15
$it \rightarrow en$	4.71 ± 0.49	4.89 ± 0.32
$en \rightarrow it$	4.72 ± 0.48	4.97 ± 0.12
$es \rightarrow it$	4.51 ± 0.49	4.91 ± 0.22
$it \rightarrow es$	4.63 ± 0.49	4.93 ± 0.20

Table 1: Average scores (0 to 5) and standard deviation of the human evaluation on *Accuracy* and *Fluency*

```
<lex quality="gold" lid="Id1" lang="en">
    Themisthocles was born in Turkey and
   he was a citizen of the Classical Athens.
    Themistocles's spouse is Archippe.
</lex>
<lex quality="silver" lid="Id2" lang="es">
    Temístocles nació en Turquía y fue
    ciudadano de la Atenas clásica.
   El cónyuge de Temístocles es Arquippe.
</lex>
<lex quality="silver" lid="Id3" lang="it">
   Temistocle è nato in Turchia ed era un
    cittadino dell'Atene classica.
   La consorte di Temistocle era Archippe.
</lex>
```

To ensure the quality of the *silver* dataset (produced during the MT and APE phases), we manually evaluated sentences that were assigned a xCOMET sentence-level score below *good* (i.e., rated as *moderate* or *weak*). We sampled 50 verbalizations for each source–target language pair and involved eight volunteer native speakers with high proficiency in at least one of the considered language pairs (e.g., English and Italian). Each source-target language pair was annotated by two different annotators following a widely adopted taxonomy for the evaluation of Accuracy and Fluency (Lommel et al., 2014) (evaluation guidelines in Appendix F.2). In addition, we asked the annotators to propose a corrected version of the automatically translated verbalization, increasing the number of gold references. In Table 1, the scores of the manual evaluation are reported. The annotators rated the sentences very highly across all language pairs, with Accuracy and Fluency consistently scoring above 4 on a scale from 0 to 5 (Inter-annotator agreement is reported in Appendix F.3). The high scores attributed to the silver standard suggest that the translated instances were considered reliable by evaluators, confirming the high quality of data, especially considering that the evaluation was conducted on inputs that received lower scores from xCOMET.

After the translation and APE phases, we obtain a dataset of $1,845$ data units, consisting of $1,120$ silver and 725 gold instances. The full distribution by language is reported in Table 5 in Appendix E.

4 Methodology

To study multilingual RDF-to-Text verbalization, we leverage three families of open state-of-the-art multilingual instruction-tuned models, each evaluated at two comparable sizes: Llama-3.2-3B-Instruct, Llama-3.1-8B-Instruct, Qwen2.5-3B-Instruct, Qwen2.5-7B-Instruct, Gemma-3-4B-IT, and Gemma-3-12B-IT. We focus on these model families because they are multilingual, widely used, and well established in the literature, openly available, and offered in moderate parameter ranges that enable efficient and reproducible experimentation.

4.1 Experimental Setup

Given a set of RDF-style triples (s, p, o), the system produces a single-paragraph natural language description in English, Spanish, or Italian. We run experiments on two datasets: (i) WebNLG, used as a baseline benchmark, and (ii) TailNLG, used to analyze performance on data units containing long-tail entities. Since our goal is to study biases and examine how LLMs handle rare entities, we deliberately adopt a zero-shot setting to minimize external influence on the model's outputs and to better observe its default behavior in isolation. Prompts are written in English, Spanish, or Italian depending on the target verbalization. Full prompts are available in Appendix H.

For each input we generate three candidate verbalizations to capture decoding variability. We use stochastic decoding with temperature sampling (temperature $= 0.7$), which increases output diversity compared to greedy decoding while keeping generations coherent.

4.2 Evaluation

In evaluating the results, we compare multiple metrics designed to capture three characteristics of the generated output: (i) semantic similarity (embedding-based metrics), (ii) textual similarity (overlap-based metrics), and (iii) model uncertainty in text generation[6]. We use BERTScore (Zhang et al., 2019) to evaluate semantic similarity, computing pairwise cosine similarity between candidate and reference texts with contextual embeddings. Specifically, we adopt the rescaled-with-baseline variant (BERTScore$_r$), using a multilingual BERT model[7] to ensure comparability across

[6]Evaluations are performed using Hugging Face packages
[7]https://huggingface.co/google-bert/bert-base-multilingual-cased

languages. To evaluate textual similarity, we adopt: BLEU (Papineni et al., 2002), chrF++ (Popović, 2017), and three variants of ROUGE (Lin, 2004): $ROUGE_1$, which measures unigram (word-level) overlap, $ROUGE_2$, which measures bigram overlap and $ROUGE_L$, which captures sentence-level structural similarity by identifying the longest common subsequences.To quantify model uncertainty, we computed the average PPL by feeding each generated verbalization back into the same model that produced it and measuring the exponential of the resulting cross-entropy loss. Lower PPL values are interpreted as a proxy for higher model confidence in text generation. To assess the statistical significance of each metric, we applied the Wilcoxon–Mann–Whitney test (Fagerland and Sandvik, 2009) on triple verbalization of the same model on different populations (e.g. long-tail *vs* head for all the languages): if the test demonstrated statistically significant variation, we marked the score with an asterisk implying a better performance. To ensure robustness and reduce the effect of sampling variability, we repeated the test ten times on random samples of 500 predictions and averaged the results across iterations..

5 Results

In this section, we present results comparing long-tail and head entities, as well as across languages (RQ1). We also examine the model uncertainty on verbalizing rare and well-known entities by comparing PPL scores on TailNLG and WebNLG (RQ2).

5.1 [RQ1] Do LLMs perform differently in the verbalization of rare entities and does language impact the verbalization output?

Table 2 reports the scores for triples containing long-tail and head entities, broken down by language, for the bigger size models. Complete results are provided in Appendix I. In all cases, $BERTScore_r$ scores higher on head triples. Conversely, results from chrF++, which correlates with morphological accuracy, and ROUGE are less consistent: in some cases they score higher on long-tail triples, while in others on head triples. The average PPL instead provides more homogeneous and clear insights, particularly for Qwen and Llama: average scores are higher for long-tail entities, suggesting that these entities present a more challenging generation task for models. Investigating the influence of language, we observe that across all settings,

Type	Lang	B_r ↑	BLEU ↑	chrF ↑	R1 ↑	R2 ↑	RL ↑	PPL ↓
Llama-8B								
Long	en	0.61	0.25	60.39	0.66	0.43	0.56	93.36
	es	0.58	0.17	53.30	0.58	0.35	0.49	222.16
	it	0.46	0.13	47.73*	0.49*	0.29*	0.43*	88.06
Head	en	0.65*	0.26	61.20	0.67	0.44	0.57	59.03*
	es	0.60	0.19	53.23	0.58	0.36	0.48	94.52*
	it	0.45	0.13	45.58	0.46	0.27	0.40	68.84*
Gemma-12B								
Long	en	0.63	0.26	61.65	0.67	0.44	0.58	330.02*
	es	0.62	0.21	58.80	0.62	0.40	0.53	611.42*
	it	0.61	0.24*	60.68*	0.64*	0.42*	0.55*	425.27*
Head	en	0.66*	0.27	61.72	0.68	0.45	0.57	397.46
	es	0.64*	0.22	57.51	0.61	0.39	0.51	926.65
	it	0.61	0.20	56.97	0.59	0.38	0.50	1068.31
Qwen-7B								
Long	en	0.65	0.29	62.47	0.69	0.46	0.58	134.13
	es	0.60	0.16	53.42	0.57	0.34	0.49	181.75
	it	0.44	0.11	46.82	0.47	0.26	0.41	100.26
Head	en	0.69*	0.29	63.06	0.70	0.47	0.59	98.05*
	es	0.61	0.16	51.83	0.55	0.32	0.46	102.36*
	it	0.46	0.10	45.35	0.45	0.25	0.39	63.48*

Table 2: Results of long vs head entities verbalizations by data type and language. Metrics include BERTScore Rescaled (B_r), BLEU, chrF++ (chrF), $ROUGE_1$ (R1) , $ROUGE_2$ (R2), $ROUGE_L$ (RL) and Perplexity (PPL).

performance is consistently higher for English, suggesting that models handle English content more effectively. Moreover, for Llama and Qwen, Italian scores are consistently lower than those for English and Spanish, indicating a possible language bias toward more widely represented languages.

Since the comparison between long-tail and head entities does not always yield univocal results, we further investigate performance across different entity popularity groups, defined according to the number of claims associated with each entity (e.g., 1–5, 6–10). Results for ROUGE-L, BERTScore Rescaled and PPL are reported in Figure 2; full results are provided in Figure 5 in Appendix I. For all metrics except ROUGE-L, results show that data units containing entities with very few claims (1–5) achieve the lowest scores. Performance then increases, reaching its highest values for head entities with 51+ claims. This suggests that entities associated with more claims may be easier for the model to verbalize, which is consistent with our hypothesis. ROUGE-L higher scores in the verbalization of long-tail entities could depend on the sequential ordering of tokens: when the model expresses the same facts using a different ordering, or follows the order of the input triples rather than the ordering adopted in the reference, semantic metrics improve, while ROUGE-L remains stable or even decreases. For example, given the triples Narita International Airport | located in time zone|

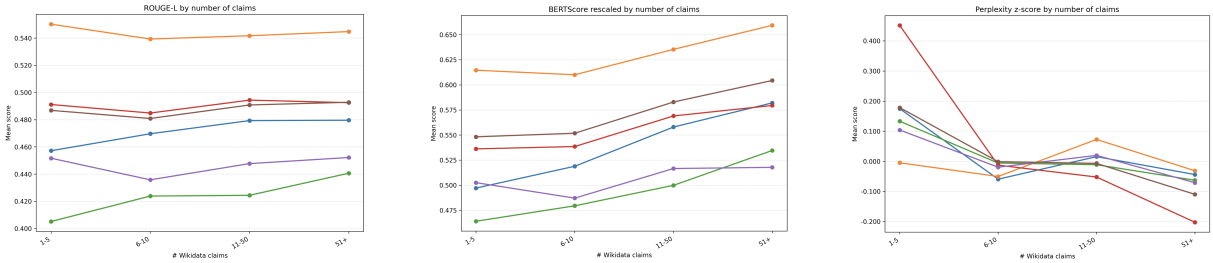

Figure 2: Average ROUGE-L, BERTScore Rescaled, and Perplexity values split by number of claims. On the x-axis, entities are grouped by number of claims, and on the y-axis, the average metric scores are reported
Gemma-3-4B, Gemma-3-12B, Llama-3.2-3B, Llama-3.1-8B, Qwen-2.5-3B, Qwen-2.5-7B

Asia/Tokyo, Narita International Airport | named after | Tokyo, the model's output is: "Narita International Airport is located in the time zone Asia/Tokyo. Narita International Airport is named after Tokyo." In contrast, the reference states: "Narita International Airport is named after Tokyo, and its time zone is Asia/Tokyo." Here, the difference in ordering leads to a lower ROUGE-L score for the prediction. Insights from PPL further strengthen these results: average perplexity is higher for data units containing entities associated with a lower number of claims (indicating higher uncertainty) and decreases as the number of claims increases. Finally to better understand whether the presence of silver standard data influenced performance, we also split the results by gold *vs.* silver (Tables 9 and 10 in the Appendix). Overall, scores are higher for the silver standard, confirming the findings of (Panickssery et al., 2024) that models are better aligned with machine-translated inputs. When results are further divided into long-tail and head entities (Table 10), statistically significant differences show higher Silver scores for head entities, suggesting that machine translation and post-editing models may also be biased against less frequent entities.

5.2 [RQ2] How does model performance differ when comparing the TailNLG benchmark with WebNLG?

Table 3 reports the average performance of each model on the WebNLG test set and on long-tail triples from TailNLG. All statistically significant results indicate that embedding- and overlap-based metrics achieve higher scores on WebNLG than in the long-tail setting. This trend suggests that models generate more accurate and lexically aligned verbalizations for well-known entities. Within each model, PPL is even lower for WebNLG verbalizations, with statistical significance, supporting the hypothesis that WebNLG entities and triples are more familiar to the models, potentially due to exposure during pre-training. In addition, since models tend to achieve higher scores when synthetic data are used as ground truth (Sarti and Nissim, 2024), our findings are further reinforced. Even when comparing model performance on the English WebNLG test set (entirely gold) with the long-tail portion of TailNLG (containing both gold and silver references), models still achieve higher scores on WebNLG. For further details, we refer the reader to Appendix I.

Type	$BERT_r$ ↑	BLEU ↑	chrF ↑	R1 ↑	R2 ↑	RL ↑	PPL ↓
Llama-8B							
Long-tail	0.55	0.18	53.81	0.58	0.36	0.50	134.53
WebNLG	0.53	0.21*	53.77	0.65*	0.43*	0.53*	68.23*
Gemma-12B							
Long-tail	0.62	0.24	60.38	0.64	0.42	0.55	455.57
WebNLG	0.65*	0.30*	63.63*	0.73*	0.50*	0.59*	290.10*
Qwen-7B							
Long-tail	0.56	0.19	54.24	0.58	0.35	0.49	138.71
WebNLG	0.57	0.24*	56.49	0.66*	0.44*	0.54*	95.54*

Table 3: Performance comparison of the average scores for models on long-tail triples from the TailNLG and WebNLG test sets.

6 Discussion

Overall, Results show that, across all models, the verbalization of triples that contain long-tail entities is more challenging (RQ1). This suggests that LLMs could suffer from entity bias, with widely represented entities being processed more easily and producing more accurate outputs.

Long-tail vs head The fact that overlap-based metrics can be higher for data units involving long-tail entities suggests that verbalizations of head entities often include additional information beyond the reference, likely due to prior knowledge

associated with these entities, leading to more verbose outputs. For example, the Italian verbalization of the data unit (Karl Marx | spouse | Jenny von Westphalen; Karl Marx | country of citizenship | statelessness) is: "Karl Marx è stato apatico alla cittadinanza, preferendo lo status di apolidia. Karl Marx ebbe come consorte Jenny von Westphalen" (English translation: Karl Marx was apathetic toward citizenship, preferring statelessness. Karl Marx's spouse was Jenny von Westphalen), which is more verbose and deviates from the reference. A similar example is the following: the gold standard (Italian) is 'Google, che prende il nome da Barney Google, è stata fondata da Larry Page' ("Google, which is named after Barney Google, was founded by Larry Page"), while the generation is 'Google è stata fondata da Larry Page e prende il nome dal personaggio Barney Google, famoso in alcuni giochi da tavolo degli anni '30' ("Google was founded by Larry Page and is named after the character Barney Google, known for board games from the 1930s."). The model adds extra information about Barney Google, describing him as associated with board games from the 1930s.

Difference in languages The results concerning language strongly confirm the hypothesis that models also suffer from language bias, and that verbalizations in English are more accurate. These results may also be influenced by the fact that both Italian and Spanish have richer morphologies than English. For example, the Italian verbalization of the triples Amanda Ammann | country of citizenship | Switzerland; Amanda Ammann | occupation | model results in "Amanda Ammann è una cittadina svizzera e lavora come modello" (English translation: Amanda Ammann is a Swiss citizen and works as a model), which exhibits a gender disagreement between the first part of the sentence ("cittadina svizzera") and the final noun ("modello"). Italian output generally tend to be less accurate and suffer from some additions and omissions. Additionally, for the Italian language, models sometimes fail to generate natural language sentences and instead simply reproduce the input triples (e.g., subject: 'Serra dos Ancares', predicate: 'located in the administrative territorial entity', object: 'León Province').

Perplexity as uncertainty proxy Another important finding emerging from the results is the high variability across evaluation metrics. In particular, reference-based metrics do not always seem able to detect the presence of bias or provide reliable scores, as becomes evident from the comparison with PPL, which yields more consistent results. The use of PPL as a proxy for uncertainty confirms our hypothesis that models perform better on benchmarks such as WebNLG than on TailNLG (RQ2), likely because they were exposed during pre-training to information related to WebNLG or its entities. This finding is supported, with statistical significance, by both reference-based evaluation metrics and PPL.

Is all that glitters gold? The results, broken down by silver and gold, also have practical implications for downstream applications that rely on verbalization, showing that in pipelines where verbalized outputs are used as input to LLMs, silver-standard data may suffice, as models appear better aligned with machine-translated or automatically generated text. This suggests that silver data could provide usable outputs while reducing the need for fully curated gold-standard annotations, although further research is needed to evaluate their impact on model performance and reliability.

7 Conclusion and Future Work

In this work, we conduct an in-depth evaluation of three state-of-the-art language model families in depth for the verbalization of long-tail entities in a multilingual setting. To this end, we introduced TailNLG, the first benchmark of triple–verbalization pairs covering entities that range from rare to well-represented. Our results indicate that the evaluated models exhibit both language- and entity-specific biases, and perform better on established benchmarks such as WebNLG than on TailNLG. In line with prior work, our findings suggest that current models still face challenges when handling long-tail entities, revealing limitations that make tasks involving less-represented entities more difficult and less reliable.

As future work, we plan first to test other prompting methodologies, such as few-shot and chain-of-thought, to investigate if and to what extent these approaches influence the outputs. We then plan to extend the TailNLG benchmark by adding more languages and instances. We also aim to refine the dataset by incorporating culturally specific examples to investigate potential model biases. Additionally, we plan to create a gold standard that reflects diverse annotator perspectives, enabling the study of variation in human judgments in the verbalization task.

Limitations

We acknowledge that building a new high-quality benchmark is challenging, as it requires achieving an appropriate balance between human annotation and automatic processing steps. In this work, we consistently paired the automatic phase with human assessment in an effort to ensure the highest possible quality. Nevertheless, the presence of *silver* verbalizations in the benchmark constitutes a limitation, since machine translation tools, although reliable, may not fully align with human judgments.

Second, by leveraging a well-known resource such as Wikidata, the variability of available claims is limited. As a result, models may already be familiar with many of the relations, which also recur across different data units.

Finally, in defining long-tail entities, we adopted a simplified approach for experimental purposes, considering only a quantitative parameter (number of claims). However, we are aware that entity underrepresentation is a multifaceted issue that involves multiple dimensions, both vertical (numerosity) and horizontal (cultural aspects). In future work, it would be valuable to consider additional perspectives and to define long-tail entities using a more comprehensive characterization.

Ethical Considerations

Our work relies on knowledge extracted from Wikidata, whose community of contributors has recognized the underrepresentation of non-Western and women contributors. This has an impact on how knowledge is shaped and which entities suffer underrepresentation. Adopting a purely statistical approach to the definition of *long-tail* entities might have resulted in overlooking these cultural factors. Nevertheless, this approach was preferred to the adoption of theoretically grounded models of underrepresentation, as the latter could have introduced additional research bias in corpus creation.

All annotations have been performed voluntarily by the members of our research group without relying on external annotation platforms.

During the preparation of this work, the authors used AI tools in order to perform grammar and spelling checks. After using these tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

Julia Adams, Hannah Brückner, and Cambria Naslund. 2019. Who counts as a notable sociologist on wikipedia? gender, race, and the "professor test". *Socius*, 5:2378023118823946.

Kancharla Aditya Hari, Bhavyajeet Singh, Anubhav Sharma, and Vasudeva Varma. 2023. WebNLG challenge 2023: Domain adaptive machine translation for low-resource multilingual RDF-to-text generation (WebNLG 2023). In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 93–94, Prague, Czech Republic. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

Marta Boscariol, Luana Bulla, Lia Draetta, Beatrice Fiumanò, Emanuele Lenzi, and Leonardo Piano. 2024. Evaluation of llms on long-tail entity linking in historical documents (short paper). In *EKAW (Companion)*.

Thiago Castro Ferreira, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina, editors. 2020. *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*. Association for Computational Linguistics, Dublin, Ireland (Virtual).

Liam Cripwell, Anya Belz, Claire Gardent, Albert Gatt, Claudia Borg, Marthese Borg, John Judge, Michela Lorandi, Anna Nikiforovskaya, and William Soto Martinez. 2023. The 2023 webnlg shared task on low resource languages. overview and evaluation results (webnlg 2023). In *Association for Computational Linguistics*, page 55–66.

Bayu Distiawan, Jianzhong Qi, Rui Zhang, and Wei Wang. 2018. Gtr-lstm: A triple encoder for sentence generation from rdf data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1627–1637.

Lia Draetta, Marco Antonio Stranisci, Flaviana Corallo, Pier Felice Balestrucci, Michael Oliverio, Rossana Damiano, and Alessandro Mazzei. 2025. Beyond the metrics: an investigation into the reliability of evaluation metrics for domain specific graph-based question answering.

Daniel Duma and Ewan Klein. 2013. Generating natural language from linked data: Unsupervised template extraction. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*, pages 83–94. ASSOC COMPUTATIONAL LINGUISTICS-ACL.

Morten W Fagerland and Leiv Sandvik. 2009. The wilcoxon–mann–whitney test under scrutiny. *Statistics in medicine*, 28(10):1487–1497.

Thiago Castro Ferreira, Diego Moussallem, Emiel Krahmer, and Sander Wubben. 2018. Enriching the webnlg corpus. In *Association for Computational Linguistics*, page 171–176.

Hanning Gao, Lingfei Wu, Po Hu, Zhihua Wei, Fangli Xu, and Bo Long. 2021. Rdf-to-text generation with reinforcement learning based graph-augmented structural neural encoders. *CoRR*.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017a. Creating training corpora for nlg micro-planning. In *55th Annual Meeting of the Association for Computational Linguistics, ACL 2017*, pages 179–188. Association for Computational Linguistics (ACL).

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017b. The webnlg challenge: Generating text from rdf data. In *10th International Conference on Natural Language Generation*, pages 124–133. ACL Anthology.

Arianna Graciotti, Leonardo Piano, Nicolas Lazzari, Enrico Daga, Rocco Tripodi, Valentina Presutti, and Livio Pompianu. 2025. Ke-mhisto: Towards a multilingual historical knowledge extraction benchmark for addressing the long-tail problem. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20316–20339.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Aidan Hogan, Xin Luna Dong, Denny Vrandečić, and Gerhard Weikum. 2025. Large language models, knowledge graphs and search engines: A crossroads for answering users' questions. *arXiv preprint arXiv:2501.06699*.

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Ming Jiang and Mansi Joshi. 2024. CPopQA: Ranking cultural concept popularity by LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 615–630, Mexico City, Mexico. Association for Computational Linguistics.

Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*, 3rd edition. Online manuscript released August 24, 2025.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, pages 15696–15707.

Rohan Kumar, Youngmin Kim, Sunitha Ravi, Haitian Sun, Christos Faloutsos, Ruslan Salakhutdinov, and Minji Yoon. 2024. Automatic question-answer generation for long-tail knowledge. *arXiv preprint arXiv:2403.01382*.

Huihan Li, Yuting Ning, Zeyi Liao, Siyuan Wang, Xiang Lorraine Li, Ximing Lu, Wenting Zhao, Faeze Brahman, Yejin Choi, and Xiang Ren. 2024. In search of the long-tail: Systematic generation of long-tail inferential knowledge via logical rule guided search. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2348–2370, Miami, Florida, USA. Association for Computational Linguistics.

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

Arle Lommel, Aljoscha Burchardt, Maja Popović, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit. 2014. Using a new analytic measure for the annotation and analysis of MT errors on real data. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pages 165–172, Dubrovnik, Croatia. European Association for Machine Translation.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822.

Sebastien Montella, Betty Fabre, Tanguy Urvoy, Johannes Heinecke, and Lina M Rojas Barahona. 2020. Denoising pre-training and data augmentation strategies for enhanced rdf verbalization with transformers. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 89–99.

Michael Oliverio, Pier Felice Balestrucci, Alessandro Mazzei, and Valerio Basile. 2024. Dipinfo-unito at the gem'24 data-to-text task: Augmenting llms with the split-generate-aggregate pipeline. In *Proceedings of the 17th International Natural Language Generation Conference: Generation Challenges*, pages 59–65. Association for Computational Linguistics.

Michael Oliverio, Pier Felice Balestrucci, Alessandro Mazzei, and Valerio Basile. 2025. Webnlg-it: Construction of an aligned rdf-italian corpus through machine translation techniques. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12073–12083.

Riccardo Orlando, Pere-Lluís Huguet Cabot, Edoardo Barba, and Roberto Navigli. 2024. Relik: Retrieve and link, fast and accurate entity linking and relation extraction on an academic budget. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14114–14132.

Arjun Panickssery, Samuel Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Laura Perez-Beltrachini, Rania Sayed, and Claire Gardent. 2016. Building RDF content for data-to-text generation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1493–1502, Osaka, Japan. The COLING 2016 Organizing Committee.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Virginia Ramón-Ferrer, Carlos Badenes-Olmedo, and Oscar Corcho. 2026. Spanish triple-to-text benchmark on low-resource large language models. *Semantic Web*, 17(3):22104968261440701.

Gabriele Sarti and Malvina Nissim. 2024. IT5: Text-to-text pretraining for Italian language understanding and generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9422–9433, Torino, Italia. ELRA and ICCL.

Anastasia Shimorina, Elena Khasanova, and Claire Gardent. 2019. Creating a corpus for russian data-to-text generation using neural machine translation and post-editing. In *Association for Computational Linguistics*, pages 44–49.

Marco Antonio Stranisci, Eleonora Bernasconi, Viviana Patti, Stefano Ferilli, Miguel Ceriani, and Rossana Damiano. 2023. The world literature knowledge graph. In *International Semantic Web Conference*, pages 435–452. Springer.

Kai Sun, Yifan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2024. Head-to-tail: How knowledgeable are large language models (LLMs)? A.K.A. will LLMs replace knowledge graphs? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 311–325, Mexico City, Mexico. Association for Computational Linguistics.

Marcos V Treviso, Nuno M Guerreiro, Sweta Agrawal, Ricardo Rei, José Pombal, Tania Vaz, Helena Wu, Beatriz Silva, Daan Van Stigt, and Andre Martins. 2024. xTower: A multilingual LLM for explaining and correcting translation errors. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15222–15239, Miami, Florida, USA. Association for Computational Linguistics.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Fan Zhang, Meishan Zhang, Shuang Liu, Yueheng Sun, and Nan Duan. 2023. Enhancing rdf verbalization with descriptive and relational knowledge. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(6):1–18.

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

A Entity Perplexity Analysis

To compute PPL as a proxy for model uncertainty, we evaluated the six language models described in Section 4. This allows us to assess whether the PPL gap between head and long-tail entities is consistent across different model families. For each model, we tokenize the entity string directly using its specific tokenizer. PPL is calculated as the exponentiated average negative log-likelihood of the sequence. We performed a statistical comparison between head and long-tail distributions using the two-sided Mann-Whitney U test. Results are reported in Figure 3 using median and interquartile range (IQR, Q1–Q3) and are visualized on a $\log_{10}$ scale to improve readability. In all settings, PPL values for long-tail entities are higher with statistical significance in most cases, thus validating our choice of using the Pareto cut-off for the selection of long-tail and top-head entities.

B Triples configuration

In line with other resources in the field of NLG, we extracted entities using different configurations. Some examples are shown in Figure 4.

C Verbalization guidelines

The annotators were asked to verbalize the data-units following guidelines:

C.1 Verbalization

- Verbalize the triples as naturally as possible, aiming to create grammatically correct and fluent sentences. You may choose whether or

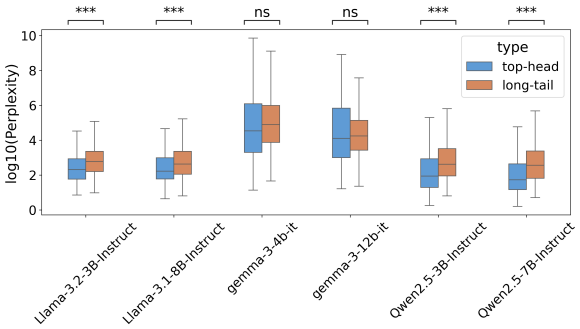

Figure 3: PPL distributions for head and long-tail entities across models (log10 scale). Boxes represent the interquartile range (IQR, Q1–Q3) with the median shown as a line inside the box. Statistical significance was assessed using the *Mann-Whitney U test*: *** $p < 0.001$, ns = not significant.

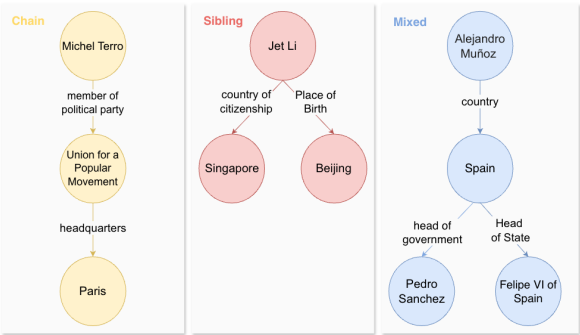

Figure 4: Examples of Chain, sibling and mixed triples configuration

not to follow the original order of the triples. Please indicate this in the "Order preserved" column by writing YES if the triple is verbalized in the same order, or NO if it is not.

- You are free to break down the information conveyed by each set of triples in one or more sentences according to your preferences.
- If a triple makes no sense (e.g., "Black Racer", "different from", "Black Racer"), you may choose not to verbalize it. In this case, write NA in the annotation column.
- If a triple is very difficult or uncertain to interpret, you may propose a verbalization and leave comments in the Notes column. A colleague will review uncertain verbalizations later.
- If you have the necessity to check the meaning or the translation of certain entities of claim you can use WikiData `https://www.wikidata.org/wiki/Wikidata:Main_Page`.
- Named entities
  - For named entities, if an equivalent exists in Italian or Spanish, please use the correct translation in your verbalization.
  - If you have doubts about the translation of a claim, you can find the corresponding version in Spanish or Italian on Wikidata.
  - If you cannot find an equivalent in Italian or Spanish, do not translate it. Instead, flag the triple as problematic in the "Notes" column.

C.2 Validation

In this phase, the annotators will cross-check and review colleagues' annotations, evaluating whether each verbalization is acceptable by answering the following questions:

- Does the text sound fluent and natural?
- Does the text include all and only the information from the data?
- Is the text written in good language (no spelling or grammatical mistakes)?

D Data Unit distribution per Category

Category	Count
Politician	44
Airport	43
Model	27
SportsTeam	30
Building	44
Company	43
Artist	37
City	40
MeanOfTransportation	12
Food	34
CelestialBody	14
Astronaut	31
Monument	42
ComicsCharacter	40
River	30
University	42
WrittenWork	19
Athlete	43
Type	**Count**
Long-tail	366
Head	249
Total entries	**615**

Table 4: Distribution of data entries across categories and entity types per language

Language	Gold	Silver	Total
English	168	447	615
Spanish	173	442	615
Italian	384	231	615
Total	725	1,120	1,845

Table 5: Distribution of gold and silver verbalizations by language in TailNLG.

E Gold and Silver Entities Distribution

F Translation

F.1 Error correction

You are provided with a Source, Translation, Translation quality analysis, and Translation quality score (weak, moderate, good, excellent). The Translation quality analysis contain a translation with marked error spans with different levels of severity (minor, major or critical). Additionally, we may provide a **reference translation**. Given this information, generate an explanation for each error and a fully correct translation.
Italian source: Osvaldo il congilio è un personaggio dell'opera Lupo Alberto
English translation: Osvaldo the congil is a character from the opera Wolf Albert.
Translation quality analysis: Osvaldo the con<error severity="minor">gil</error> is a character from the opera<error severity="minor">Wolf Albert</error>.
Translation quality score: good

The corresponding output is:

Explanation for error: The term "congilio" is a typographical error and should be corrected to "coniglio," which means "rabbit" in Italian. The correct translation of "Lupo Alberto" is "Wolfy Albert," as "Lupo" means "wolf" and "Alberto" is a common name in Italian. The translation also misses the gender of "Osvaldo," which is masculine, and the correct translation should reflect that. Additionally, the translation of "opera" as "opera" is correct, but the translation of "Wolf Albert" as "Wolf Albert" is incorrect because it does not convey the correct name of the work. The correct translation should reflect the proper name of the work, which is "Lupo Alberto."

F.2 Guidelines for Human MT Evaluation

Annotators were instructed to strictly follow the evaluation guidelines and, when necessary, to correct the MT output by providing a correct or more accurate version.

Annotate as **1** if the error is present and **0** if not, respecting the following taxonomy:

Accuracy

Issues related to whether the information content of the target is equivalent to the source. Subcategories:

- **Terminology:** Issues related to the use of domain-specific terms.
- **Mistranslation:** Issues related to the improper translation of content, including erroneous translation of named entities.
- **Omission:** Content present in the source is missing in the target.
- **Addition:** Content not present in the source has been added to the target.
- **Untranslated:** Text inappropriately appears in the source language.

Fluency

Issues related to the linguistic properties of the target without relation to its status as a translation. Subcategories:

- **Grammar:** Issues related to grammatical properties of the text.
- **Style:** The text shows stylistic problems.
- **Spelling:** The text is spelled incorrectly.
- **Typography:** Problems related to typographical conventions.
- **Unintelligible:** Text is garbled or otherwise unintelligible; indicates a major breakdown in fluency.

Corrected Translation

If the target translation is incorrect, you may propose a new translation in the "Corrected translation" column.

Guidelines for Annotators

- Evaluate whether translating a Named Entity (NE) makes sense in context, taking into account cultural and linguistic conventions; some NEs may be left untranslated.
- Named entities should be considered correctly translated when it is common in the target language to refer to the entity in that way; otherwise, mark them as incorrect. Examples:
  - Crist Redemptor → Cristo Redentore → Christ the Redeemer [Correct]
  - Lupo Alberto → Wolfy Albert [Incorrect]
  - Louis the Springer → Luigi il Precursore [Incorrect]

  Wrong NEs must be annotated in the column "Mistranslation".
- Assess when it is appropriate to keep a term in English (or another source language).
- Pay attention to the correct use of verb tenses.
- Ensure agreement between singular and plural.
- The translation should be read naturally in the target language.
- Ensure that the meaning is accurately preserved.
- Annotators should mark **1** for each subcategory when an issue occurs. If you label **1**, specify why in the "note" column.

F.3 Manual assessment of translation

In table 6 are reported the Spearman and Pearson correlation between the two annotators.

Source	Target	ρ	r
es	en	0.23	0.25
en	es	0.32*	0.27
it	en	0.15	0.02
en	it	0.56*	0.64
es	it	0.26	0.22
it	es	0.53*	0.49*

Table 6: Spearman (ρ) and Pearson (r) correlation coefficients across language pairs. Source and Target denote the source language and the target language into which the source is translated, respectively. Asterisks (*) denote a statistically significant difference ($p < 0.05$).

G TailNLG Benchmark

Each entry contains the following metadata labels:

- **Category**: domain/class of the subject (e.g., Artist, City, Monument).
- **Eid**: internal entry id (e.g., `Id1`).
- **Shape_type**: data Units configuration (chain, sibling, mixed) that controls structural complexity.
- **Size**: number of triples for Data Units.
- **Qid**: Wikidata Id.
- **Type**: Distinction between head and long-tail entities
- **Sub_type**: Distinction between subtypes (e.g. head entities with rare claims).
- **triplesets**: `originaltripleset` (as collected) and `modifiedtripleset` (normalized input).
- **lexicalizations**: one or more `lex` elements with attributes `quality` (gold/silver) and `lang` (en/es/it).

H Prompts

English version In English, structured data is commonly represented as triples, with the format [subject, predicate, object]. Based on these triples, generate a single-paragraph text composed of complete, grammatically correct, and natural sentences.
INSTRUCTIONS:
- Generate the text solely from the input triples
- Return the final verbalization with this format: The final verbalization is: [verbalization output]
- Insert the verbalization of the input triples within square brackets [], without adding anything else.

Spanish version En español, los datos estructurados se representan comúnmente como tríos, con el formato [sujeto, predicado, objeto]. Basándose en estos tríos, genere un texto de un solo párrafo compuesto por oraciones completas, gramaticalmente correctas y naturales.
INSTRUCCIONES:

- Genere el texto únicamente a partir de las tripletas de entrada.
- Devuelva la verbalización final con este formato: La verbalización final es: [salida de verbalización]
- entre corchetes [] inserte la verbalización de las tripletas de entrada, sin añadir nada más

Italian version In italiano, i dati strutturati sono comunemente rappresentati come triple, con il formato [soggetto, predicato, oggetto]. Sulla base di queste triple, genera un testo di un solo paragrafo composto da frasi complete, grammaticalmente corrette e naturali.
INSTRUZIONI:
- Genera il testo esclusivamente dalle triple in input
- Restituisci la verbalizzazione finale con questo formato: La verbalizzazione finale è: [output di verbalizzazione]
- tra le parentesi quadre [] inserisci la verbalizzazione delle triple in input, senza aggiungere altro.

I Results

Experiments were conducted on a server with two Intel Xeon Gold 5418Y CPUs (48 physical cores each, 96 threads total), approximately 504 GB of RAM, and two NVIDIA L40 GPUs with 46 GB of VRAM each. Generations are bounded by a fixed maximum length (max_new_tokens $= 256$) to prevent overly long outputs and to standardize comparisons across models and settings. Below, we report all results, including those across different numbers of claims (2), comparisons between data units containing long-tail and head entities across all models (7), full results comparing data unit types and languages (8), results comparing against silver and gold standards (9), and results comparing silver and gold standards split by entity type (long-tail vs. head) (11).

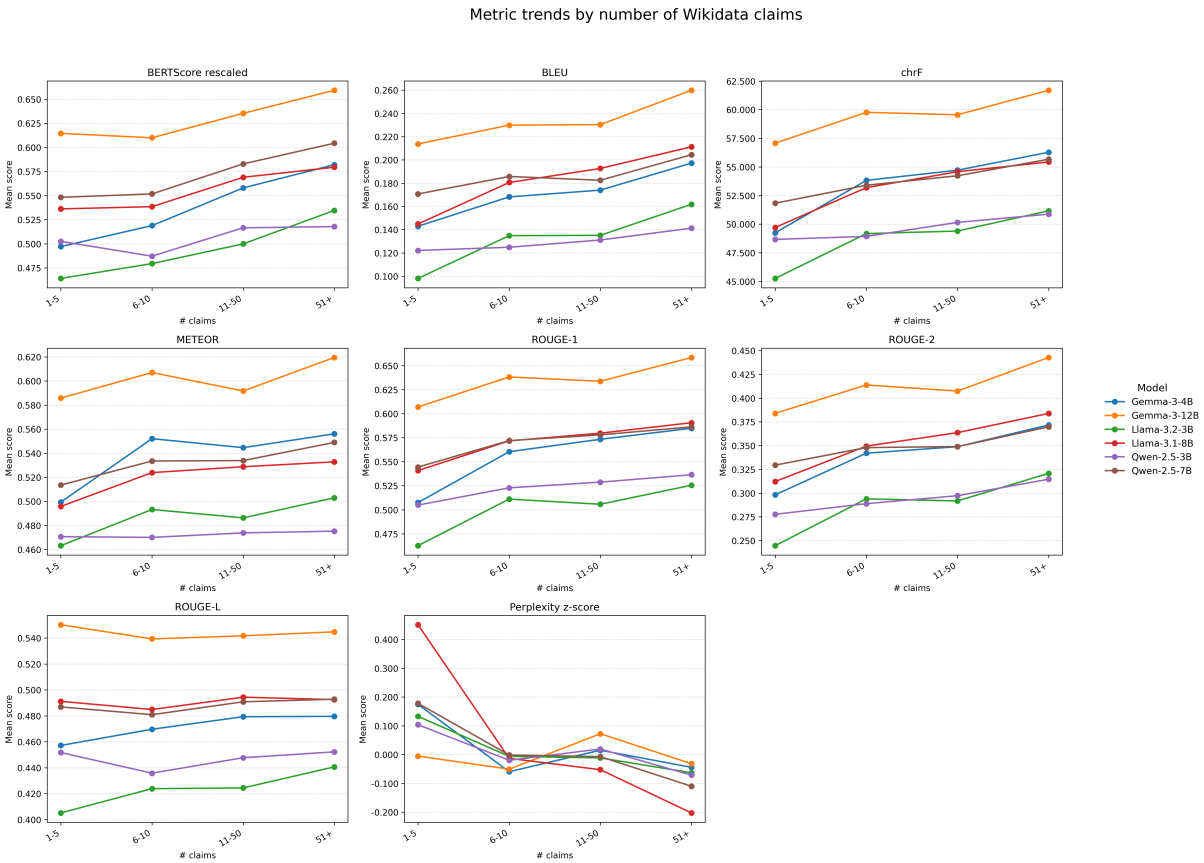

Figure 5: Average scores for BertScore rescaled, BLEU, chrf++, METEOR, ROUGE-1, ROUGE-2, ROUGE-L and Perplexity by number of claims.

Model	Type	$BERT_r$	BLEU ↑	chrF ↑	R-1 ↑	R-2 ↑	R-L ↑	PPL ↓
Qwen-3B	Long-tail	0.50	0.13	50.10	0.53	0.30	0.45	182.87
	Head	0.51	0.13	49.12	0.52	0.29	0.44	70.45*
Qwen-7B	Long-tail	0.56	0.19	54.27	0.58	0.35	0.49	138.70
	Head	0.58	0.18	53.41	0.57	0.35	0.48	87.96*
Gemma-4b	Long-tail	0.53	0.17	54.32	0.57	0.35	0.48	4724.94
	Head	0.56	0.17	53.48	0.56	0.34	0.46	3488.00*
Gemma-12b	Long-tail	0.62	0.24	60.38	0.64	0.42	0.55	455.57*
	Head	0.64	0.23	58.73	0.63	0.40	0.53	797.47
Llama-3B	Long-tail	0.49	0.13	49.75	0.51	0.30	0.43	102.40
	Head	0.50	0.14	48.19	0.49	0.29	0.42	72.55*
Llama-8B	Long-tail	0.55	0.18	53.81	0.58	0.36	0.50	134.53
	Head	0.57	0.19	53.34	0.57	0.36	0.48	74.13*

Table 7: Results of long vs head entities verbalizations by data type and language. Metrics include BERTScore Rescaled (BERTr), BLEU, chrF++ (chrF), ROUGE-1 (R-1) , ROUGE-2 (R-2), ROUGE-L (R-L) and Perplexity (PPL). Asterisks (*) denote a statistically significant difference with $p < 0.05$.

Model	Lang	$BERT_r$	BLEU ↑	chrF ↑	R-1 ↑	R-2 ↑	R-L ↑	PPL ↓
Llama-3B	en	0.58*	0.21*	56.67*	0.63*	0.40*	0.53*	73.51*
	es	0.46	0.11	46.40	0.47	0.26	0.39	114.99
	it	0.44	0.08	44.26	0.42	0.22	0.35	81.98
Llama-8B	en	0.63*	0.25*	60.72*	0.66*	0.43*	0.56*	79.37*
	es	0.59	0.18	53.27	0.58	0.35	0.49	170.14
	it	0.45	0.13	46.85	0.48	0.28	0.42	80.23
Qwen-3B	en	0.50	0.17*	52.79*	0.60*	0.36*	0.51*	84.29
	es	0.54*	0.12	49.10	0.51	0.28	0.43	227.52
	it	0.48	0.09	47.19	0.46	0.24	0.39	100.23
Qwen-7B	en	0.67*	0.29*	62.71*	0.69*	0.46*	0.59*	119.43
	es	0.60	0.16	52.78	0.56	0.33	0.48	149.16
	it	0.45	0.11	46.25	0.46	0.26	0.40	85.39*
Gemma-4b	en	0.62*	0.23*	60.32*	0.66*	0.42*	0.55*	3798.99
	es	0.51	0.14	50.76	0.53	0.31	0.44	6068.41
	it	0.50	0.14	50.85	0.50	0.30	0.43	2794.93*
Gemma-12b	en	0.64*	0.27*	61.68*	0.68*	0.44*	0.57*	357.51
	es	0.63	0.21	58.27	0.62	0.39	0.52	739.90
	it	0.61	0.22	59.17	0.62	0.40	0.53	687.37

Table 8: Average scores split by language. Metrics include BERTScore Rescaled (BERTr), BLEU, chrF++ (chrF), ROUGE-1 (R-1) , ROUGE-2 (R-2), ROUGE-L (R-L) and Perplexity (PPL). Asterisks (*) denote a statistically significant difference with p < 0.05.

Model	Quality	$BERT_r$ ↑	BLEU ↑	chrF ↑	R-1 ↑	R-2 ↑	R-L ↑	PPL ↓
Llama-3B	Gold	0.46	0.12	48.04	0.48	0.27	0.40	95.30
	Silver	0.52*	0.14	49.80	0.52*	0.30*	0.44*	86.86*
Llama-8B	Gold	0.52	0.18	52.68	0.55	0.34	0.48	90.94
	Silver	0.58*	0.19	54.22	0.59*	0.36*	0.50*	122.10*
Qwen-3B	Gold	0.50	0.13	50.62	0.53	0.30	0.44	94.12
	Silver	0.51	0.13	49.08	0.52	0.29	0.45	165.11
Qwen-7B	Gold	0.53	0.17	52.59	0.55	0.33	0.47	117.53
	Silver	0.60*	0.20	54.75	0.59*	0.36	0.50	118.35
Gemma-4b	Gold	0.52	0.17	53.67	0.55	0.34	0.46	1143.13
	Silver	0.55*	0.18	54.18	0.57	0.35	0.48	6197.47
Gemma-12b	Gold	0.61	0.23	59.88	0.63	0.41	0.54	406.83
	Silver	0.64*	0.24	59.59	0.64	0.41	0.54	715.74

Table 9: Average scores split for quality levels (Silver, Gold). Metrics include BERTScore Rescaled (BERTr), BLEU, chrF++ (chrF), ROUGE-1 (R-1) , ROUGE-2 (R-2), ROUGE-L (R-L) and Perplexity (PPL). Asterisks (*) denote a statistically significant difference with $p < 0.05$.

Type	Quality	$BERT_r$	BLEU ↑	chrF ↑	R-1 ↑	R-2 ↑	R-L ↑	PPL ↓
				Llama-3B				
Long-tail	Silver	0.50*	0.14	49.63	0.52	0.30	0.44	89.70*
	Gold	0.47	0.13	49.95	0.50	0.29	0.42	122.69
Head	Silver	0.53*	0.16*	50.04*	0.52*	0.31*	0.44*	82.59
	Gold	0.46	0.11	45.40	0.45	0.25	0.38	57.57
				Llama-8B				
Long-tail	Silver	0.56	0.18	53.64	0.58	0.35	0.50	149.01
	Gold	0.52	0.19	54.07	0.57	0.36	0.49	111.15*
Head	Silver	0.60*	0.21*	55.08*	0.60*	0.38*	0.50*	81.60
	Gold	0.52	0.16	50.76	0.53	0.32	0.45	63.08
				Gemma-4B				
Long-tail	Silver	0.54	0.17	53.96	0.57	0.35	0.48	6781.26
	Gold	0.52	0.18	54.91	0.56	0.35	0.48	1405.88*
Head	Silver	0.58*	0.18*	54.50*	0.57*	0.35*	0.48*	5319.16
	Gold	0.53	0.16	51.96	0.53	0.32	0.44	781.08
				Gemma-12B				
Long-tail	Silver	0.63*	0.24	59.82	0.64	0.41	0.55	460.31
	Gold	0.61	0.24	61.28*	0.65	0.43	0.56	447.92
Head	Silver	0.65*	0.24	59.26	0.64*	0.41*	0.54*	1100.04
	Gold	0.61	0.21	57.95	0.61	0.39	0.51	350.20
				Qwen-3B				
Long-tail	Silver	0.50	0.12	48.93	0.52	0.29	0.44	224.95
	Gold	0.50	0.15	51.91*	0.54	0.31	0.46	116.44*
Head	Silver	0.52*	0.13	49.29	0.53*	0.30*	0.45*	75.08
	Gold	0.51	0.12	48.83	0.50	0.28	0.42	63.37
				Qwen-7B				
Long-tail	Silver	0.59*	0.19	54.66	0.59	0.36	0.50	129.26*
	Gold	0.52	0.18	53.57	0.56	0.35	0.49	153.98
Head	Silver	0.62*	0.20*	54.89*	0.59*	0.36*	0.50*	101.93
	Gold	0.54	0.16	51.23	0.53	0.32	0.45	67.31

Table 10: Average scores split for quality levels (Silver, Gold) and type (long-tail vs head). Metrics include BERTScore Rescaled (BERTr), BLEU, chrF++ (chrF), ROUGE-1 (R-1) , ROUGE-2 (R-2), ROUGE-L (R-L) and Perplexity (PPL). Asterisks (*) denote a statistically significant difference with $p < 0.05$.

Type	Lang	$BERT_r$	BLEU ↑	chrF ↑	R-1 ↑	R-2 ↑	R-L ↑	PPL ↓
				Llama-3B				
Long-tail	en	0.57	0.21	56.58	0.62	0.39	0.53	86.17
	es	0.45	0.11	47.07	0.48	0.26	0.40	144.99
	it	0.44	0.09	45.61	0.44	0.23	0.37	75.80
	Total	0.49	0.13	49.75	0.51	0.30	0.43	102.32
WebNLG	en	0.56	0.23	57.17	0.71*	0.47*	0.56	53.67*
	es	0.49*	0.16*	49.20	0.58*	0.35*	0.45*	68.28*
	it	0.48*	0.13*	47.75*	0.53*	0.30*	0.42*	41.18*
	Total	0.51	0.17*	51.37	0.60	0.37	0.48*	54.38*
				Llama-8B				
Long-tail	en	0.61	0.25	60.39	0.66	0.43	0.56	93.36
	es	0.58*	0.17	53.30	0.58	0.35	0.49	222.16
	it	0.46*	0.13	47.73	0.49	0.29	0.43	88.06
	Total	0.55	0.18	53.81	0.58	0.36	0.50	134.53
WebNLG	en	0.61	0.30	65.53	0.74*	0.51*	0.59*	42.56*
	es	0.55	0.19*	52.43	0.66*	0.43*	0.53*	90.42*
	it	0.43	0.15*	45.35	0.54*	0.34*	0.46*	71.73*
	Total	0.53	0.21*	53.77	0.65*	0.43*	0.53*	68.23*
				Gemma-4B				
Long-tail	en	0.60	0.23	59.85	0.66	0.42	0.55	4886.76
	es	0.50	0.14	51.49	0.54	0.32	0.45	8074.87
	it	0.49	0.15	51.62	0.51	0.31	0.44	1213.20
	Total	0.53	0.17	54.32	0.57	0.35	0.48	4724.94
WebNLG	en	0.61	0.28*	62.43*	0.74*	0.50*	0.59*	930.34*
	es	0.52	0.20*	53.79*	0.61*	0.39*	0.48*	1493.20*
	it	0.49	0.18*	53.72	0.57*	0.36*	0.46	601.97*
	Total	0.54	0.22*	56.65*	0.64*	0.42*	0.51	1008.50*
				Gemma-12B				
Long-tail	en	0.63	0.26	61.65	0.67	0.44	0.58	330.02*
	es	0.62	0.21	58.80	0.62	0.40	0.53	611.42
	it	0.61	0.24	60.68	0.64	0.42	0.55	425.27
	Total	0.62	0.24	60.38	0.64	0.42	0.55	455.57
WebNLG	en	0.66*	0.33*	65.97	0.77*	0.53*	0.62*	333.93
	es	0.65	0.30*	62.68	0.71*	0.49*	0.57*	210.04*
	it	0.64*	0.28*	62.24*	0.70*	0.47*	0.57	326.33*
	Total	0.65*	0.30*	63.63	0.73*	0.50*	0.59*	290.10*
				Qwen-3B				
Long-tail	en	0.49	0.17	52.66	0.59	0.35	0.51	100.79
	es	0.53	0.12	49.47	0.52	0.29	0.44	325.03
	it	0.48	0.10	48.08	0.48	0.25	0.41	124.49
	Total	0.50	0.13	50.07	0.53	0.30	0.45	183.44
WebNLG	en	0.55*	0.24*	57.05	0.72*	0.47*	0.57*	51.40*
	es	0.56*	0.17*	51.54	0.63*	0.38*	0.49*	150.84*
	it	0.55*	0.15*	51.75	0.59*	0.34*	0.47*	55.81*
	Total	0.55*	0.19*	53.44	0.64*	0.40*	0.51*	86.02*
				Qwen-7B				
Long-tail	en	0.65*	0.29	62.47	0.69	0.46	0.58	134.13
	es	0.60	0.16	53.43	0.57	0.34	0.49	181.75
	it	0.44	0.11	46.82	0.47	0.26	0.41	100.26
	Total	0.56	0.19	54.24	0.58	0.35	0.49	138.71
WebNLG	en	0.68	0.34*	66.33	0.79*	0.55*	0.64*	46.52*
	es	0.60	0.24	56.19	0.66*	0.43*	0.53*	117.41*
	it	0.44	0.15	46.96	0.54*	0.33*	0.45*	122.70*
	Total	0.57	0.24	56.49	0.66*	0.44*	0.54*	95.54*

Table 11: Average scores split by language and benchmark type (TailNLG and WebNLG). Metrics include BERTScore Rescaled (BERTr), BLEU, chrF++ (chrF), ROUGE-1 (R-1) , ROUGE-2 (R-2), ROUGE-L (R-L) and Perplexity (PPL). Asterisks (*) denote a statistically significant difference with $p < 0.05$.