Who Is the Author? Investigating Gender and Geographical Biases in Scholarly Data-to-Text Generation

Lia Draetta[1,*], Michael Oliverio[1], Walid Belafrites[1], Serafeim Chatzopoulos[2], Eleni Adamidi[2], Thanasis Vergoulis[2], Miriam Baglioni[3], Alessia Bardi[3], Paolo Manghi[3], Alessandro Mazzei[1] and Rossana Damiano[1]

[1]*Department of Computer Science, University of Turin, Italy*

[2]*Athena Research Center, IMSI, Athens, Greece*

[3]*Institute of Information Science and Technologies Alessandro Faedo, National Research Council, Pisa, Italy*

Abstract

The verbalization of structured knowledge from Knowledge Graphs (KGs) into natural language is increasingly supported by Large Language Models (LLMs), improving accessibility for non-expert users in domains such as scholarly data management. However, LLMs are known to reproduce societal biases, raising concerns that automatic verbalization of researcher profiles from structured data may perpetuate gender and geographical biases, for instance by misattributing gender or mishandling non-Western names. To investigate this issue within the scholarly domain, we constructed a dedicated dataset by extracting and balancing data from the OpenAIRE knowledge graph, supplemented with SemOpenAlex, along gender and country of affiliation dimensions. We evaluated six LLMs developed across three geopolitical regions (North America, Europe, and Asia) in four languages (English, French, Italian, and Greek), assessing performance through semantic similarity, lexical overlap, and F1, precision, and recall of gender verbalization. Results show that all models exhibit a masculine bias, over-predicting male gender under uncertainty, and that gender prediction performance is consistently lower for researchers affiliated with non-Western countries. These findings highlight the risk of perpetuating gender and geographic biases in the LLM-based verbalization of scholarly knowledge graphs and underscore the need for further mitigation efforts.

Keywords

Research Domain, Knowledge Graphs, RDF-to-Text, Gender Bias

1. Introduction

The automatic verbalization of structured knowledge plays a crucial role in making Knowledge Graphs (KGs) accessible to non-expert users and in supporting retrieval-augmented generation (RAG) systems [1]. This is especially important in the research domain, where KGs are widely used to support efficient querying and exploration of large research data repositories [2, 3]. Despite their value, KGs and structured data are often difficult for non-expert audiences to access, as their formats are designed primarily for machine readability rather than human understanding. This limitation is particularly evident in research settings, where large structured repositories underpin bibliographic databases and institutional reporting for research management and university administration.

Data-Driven Storytelling: Bridging Knowledge Graphs, GenAI, and Narrative (DDS2026), in ISWC 2026 Workshops Joint Proceedings, October 25–26, 2026, Bari, Italy

*Corresponding author.

lia.draetta@unito.it (L. Draetta); michael.oliverio@unito.it (M. Oliverio); walid.belafrites@edu.unito.it (W. Belafrites); schatz@athenarc.gr (S. Chatzopoulos); eleni.adamidi@athenarc.gr (E. Adamidi); vergoulis@athenarc.gr (T. Vergoulis); miriam.baglioni@isti.cnr.it (M. Baglioni); alessia.bardi@openaire.eu (A. Bardi); paolo.manghi@openaire.eu (P. Manghi); alessandro.mazzei@unito.it (A. Mazzei); rossana.damiano@unito.it (R. Damiano)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-05

CEUR-WS.org/Vol-4313/dds-paper7.pdf

In recent years, growing attention has been devoted to the use of LLMs for the RDF-to-Text task (e.g. rewriting Resource Description Framework (RDF) data in Natural language), driven by their potential to enhance the accessibility of structured data. Prior work has primarily examined how state-of-the-art (SoTA) LLMs perform across different languages and resource settings, such as high-resource versus long-tail settings [4, 5]. At the same time, an emerging body of research has shown that LLMs can reproduce existing societal biases, thereby perpetuating harm and reinforcing the marginalization and under representation of less visible communities and individuals [6, 7]. Within the research domain, the adoption of LLMs to automatically verbalize reports and curricula from structured data may therefore risk perpetuating biases, for instance, by incorrectly referring to researchers as male when they are female or verbalizing geographical names incorrectly. While this issue is an explicit limit in Machine translation (MT) context [8, 9], it is probably also a limitation of RDF-to-Text systems, especially when multiple languages are considered. The issue, in addition, is particularly relevant in languages with rich morphology, where grammatical gender must be explicitly marked in articles and adjectives. While such models can undoubtedly represent powerful tools for data verbalization, it is essential to first investigate whether, and to what extent, they could reproduce different types of bias. In particular, while LLMs may be able to correctly infer the gender associated with Western names and produce accurate verbalization, they are likely to face greater difficulty when evaluating names of researchers from non-Western countries [10, 11].

To investigate whether LLMs exhibit biases when verbalizing data within the scholarly domain, particularly regarding researchers from diverse geographical backgrounds and with different genders, we constructed a dedicated dataset and conducted a systematic assessment across multiple models. Specifically, we extracted data from the OpenAIRE knowledge graph [2], supplemented with information from SemOpenAlex [3], two well-established open-source resource in the scholarly domain. The resulting dataset was balanced by gender and country of origin of researchers, with country classification following the Standard Country or Area Codes for Statistical Use (UN M49).[1]. We then evaluated six different LLMs developed across three geopolitical regions (North America, Europe, and Asia) to examine whether models from different regions exhibit systematic differences. We conducted the analysis in parallel across four different languages (English, French, Italian, and Greek) to investigate how models handle different languages and whether language influences the manifestation of biases. Model performance was assessed using complementary metrics capturing semantic similarity, lexical overlap with reference texts, and F1, precision, and recall of the gender in the final verbalization.

The results show that all the investigated models tend to suffer from masculine bias, overpredicting the male gender in cases of uncertainty. In addition, results show that the performance of gender prediction is consistently lower when dealing with researchers with affiliations from non-Western countries (e.g., Africa, Asia), suggesting that the models struggle to correctly identify non-Western names. These results highlight the need to address the risk of gender and geographic biases when applying LLMs to scholarly knowledge graphs and pave the way for future research on mitigating the discovered limitations.

2. Related works and backgrounds

2.1. Data-to-Text

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 [12]. 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,

[1]https://unstats.un.org/unsd/methodology/m49/

enabling fluent and information-rich verbalizations that better capture lexical and syntactic variation [13, 14]. 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 [15], 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 [16, 17].

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 [18, 19].

Recent work by [4] has investigated how different language models perform on verbalization tasks involving long-tail (i.e., rare) entities extracted from Wikidata. Their results reveal that LLMs exhibit biases when dealing with less frequent entities, suggesting that the verbalization of less widely represented entities may be less accurate and reliable.

2.2. The Scholarly Domain

In the context of scholarly domain evaluation (i.e., the assessment of scholars' careers and research impact), several recent initiatives, such as DORA[2] and CoARA[3], have emphasized the need for evaluation approaches that prioritize qualitative insights over purely quantitative metrics and citation-based indicators. These initiatives represent the institutionalization of the Open Science movement's aspirations and values into robust organizational frameworks aimed at promoting systemic and global changes that extend beyond mere openness of publications and research data [20]. In this context, the adoption of narrative curricula has emerged as a promising initiative. By moving beyond exhaustive lists of publications and achievements, narrative curricula foster a quality over quantity approach to research assessment. They reduce the reliance on journal-based metrics, acknowledge diverse and non-linear career trajectories, and allow researchers to better articulate their contributions to science, collaborative activities, and society as a whole [21]. While these approaches are promising, and SoTA LLMs could serve as useful tools to support researchers during this transition, several studies have shown that Artificial Intelligence (AI) systems often misinterpret personal names originating from non-Western cultures. Such systems may incorrectly infer gender, alter names into more familiar Western variants, or otherwise misrepresent individuals' identities [22]. These considerations are even more valuable considering that recent benchmarks reveal that LLMs perform significantly better on globally prominent cultures than on underrepresented ones [23, 24, 25], often struggling with local cultural knowledge even when queries are posed in culturally appropriate languages [26].

In this context, the present contribution aims to investigate the extent to which SoTA LLMs exhibit biases related to gender and geographical origin when verbalizing data from a well-established resource within the Open Science ecosystem, namely OpenAIRE[4].

3. Dataset creation

To investigate whether different LLMs exhibit gender and geographical biases in the scholarly domain, we leveraged two well-established scholarly knowledge graphs as data sources: the OpenAIRE Knowledge Graph [2] and SemOpenAlex [3].

Our goal was to construct a parallel dataset in four languages (English, French, Italian and Greek) consisting of RDF triples *<subject, predicate, object>* paired with their reference verbalizations. To this

[2]https://sfdora.org/
[3]https://www.coara.org/
[4]https://www.openaire.eu/

end, we designed a dataset creation pipeline inspired by previous work on data-to-text generation [27, 4]. The pipeline combines automated LLM-assisted stages with manual validation and quality assessment to ensure the correctness and consistency of the resulting dataset.

We first extracted the data from the OpenAIRE Knowledge Graph, a free and openly accessible scholarly knowledge graph that aggregates hundreds of millions of metadata records from more than 2,000 trusted research data sources. We subsequently enriched these data with complementary information unavailable in OpenAIRE, including research topics and keywords retrieved from the SemOpenAlex knowledge graph.

The resulting knowledge base, represented in RDF, was designed to be balanced with respect to both gender and country of affiliation. The following subsections describe each stage of the dataset creation pipeline in detail.

3.1. Data Extraction

As described in the previous section, our goal was to construct a dataset balanced with respect to scholars' gender and geographical affiliation by extracting information from well-established knowledge graphs. In the following subsections, we describe the data extraction and processing pipeline in detail.

OpenAIRE While gender information is not explicitly provided in the OpenAIRE Knowledge Graph, we inferred it from the authors' biographies. OpenAIRE includes biographical descriptions that are typically written by the authors themselves in the third person and often contain personal pronouns (e.g., he or she) referring to the author.

To obtain reliable gender information, we first selected only those scholar profiles that included a biography. We then inferred gender based on the personal pronouns appearing in the text, mapping he, him, and his to `male`; she and her to `female`; and they, them, and their to `other`[5].

For example, consider the following biography [6]:

```
"Name surname (born * February 199*) is a child physiotherapist and physiologist specializing
in neurodevelopment and cerebral palsy. In 2018, she received her BSc in Physical Therapy and
Rehabilitation at Bahcesehir University, Istanbul. In 2021, she received her associate degree in Health
Institutions Management, and in 2022 she graduated with an MSc in Physiology, studying neonatal hypoxic
ischemic encephalopathy at Acibadem University, Istanbul."
```

Based on the feminine pronouns (she, her) appearing in the biography, we assigned the scholar to the `female` gender category.

For each scholar, the OpenAIRE knowledge graph also provides an affiliation together with the corresponding country. To obtain a more homogeneous set of geographic categories suitable for analysis, we mapped each country to its corresponding region according to the UN M49 standard, a widely adopted classification scheme for statistical analyses. This mapping reduces geographical granularity while ensuring consistency and comparability across countries.

At this step, we obtain a dataset that, for each scholar, contains a gender, an affiliation, the relative country, and the relative UN M49 mapping in the format *<subject, predicate, object>* as following:

```
{
    "subject": "Name Surname",
    "predicate": "hasID",
    "object": "researcher_0134"
},
{
    "subject": "Name Surname",
```

[5]The cases categorized as "others" were ultimately excluded from the analysis due to severe numerical imbalance.
[6]The reported biography is anonymized for privacy reason

```
    "predicate": "hasGender",
    "object": "Female"
},
{
    "subject": "Name Surname",
    "predicate": "hasAffiliation",
    "object": "L. N. Gumilyov Eurasian National University"
},
{
    "subject": "L. N. Gumilyov Eurasian National University",
    "predicate": "location",
    "object": "Central Asia"
}
```


SemOpenAlex While the data obtained from OpenAIRE were limited to researchers' gender and affiliation, we aimed to enrich this information by integrating it with data from other trustworthy sources. The objective was twofold: first, to construct a dataset containing a sufficiently large number of triples to support meaningful verbalization; and second, to ensure an adequate degree of variation in the predicates associated with set of triples. To achieve this, we used the unique ORCID identifier of each researcher to retrieve additional information through the SemOpenAlex API, thereby enriching the data units. Specifically, we extracted the researcher's h-index, total citation count, total number of publications, research topics, relevant keywords, the three most cited publications together with their citation counts and publication years, and the co-authors with whom the researcher has the highest number of shared publications. An example of set of triples is reported in appendix A.

After these phases, we obtain a dataset comprising 500 sets of triples, balanced by gender and geographical provenance. Full statistics of the dataset are reported in Appendix A.1 Table 3.

3.2. Verbalization

To investigate whether RDF verbalizations differ across languages and whether different LLMs exhibit distinct types of bias when generating text in different languages, we require a parallel dataset consisting of RDF triples and their corresponding reference verbalizations in four languages: English, Italian, French, and Greek.

To create a reliable gold standard while avoiding the cost of fully manual annotation, we followed the methodology proposed by [4]. Specifically, the gold-standard construction consisted of two stages: (i) manual verbalization and (ii) cross-translation. In both stages, we adopted a hybrid workflow that combines LLM-assisted generation with human validation and revision.

Manual verbalization To build a full multilingual benchmark, we produced <Data_Unit, Verbalization> pairs for every data unit (set of triples) in every language. The manual verbalization process was carried out by partitioning the complete dataset of 500 data units into three subsets: two subsets containing 166 data units each and one subset containing 168 data units. Three native speakers [7] of English, French, and Italian verbalized one subset each in their respective native language. Annotators were instructed to verbalize the input triples as naturally as possible, producing fluent and grammatically correct sentences while adhering to guidelines designed to promote naturalness and ensure consistency across the dataset. The complete annotation guidelines, together with example verbalizations, are provided in Appendix A.2. Annotators were allowed to skip a data unit if they judged it to be incorrect or incomplete. For the Greek language, due to the lack of sufficiently large team of native speakers, we decided to proceed in a slightly different way: we skipped the part of manual verbalization and directly translated the input into the Greek language, checking the verbalization afterwards as described in

[7]The native speakers were two researchers and one master student, all of them experts in the field of Natural Language Generation.

the next paragraph. After this phase we obtained a set of *<RDF triple set, verbalization>* pair in three different languages.

Translation The manual verbalization phase produced a total of 500 gold verbalizations (166 in English, 166 in Italian, and 168 in French). To create a fully parallel dataset we then adopted an approach based on Machine Translation (MT), Automatic Error Detection (AED) and human-check, as proposed by Aditya Hari et al. [28]. Drawing on earlier studies on cross-lingual extensions of RDF-to-Text datasets [29, 30], we leveraged DeepL for MT.[8] For each manually created verbalization in a source language (A) corresponding translations were automatically generated in the other three target languages (B, C and D), resulting in a total of 1,500 automatically generated translations. To mitigate known issues in MT, such as omissions and hallucinations [31], we introduced an AED phase. To identify errors in the translations, we leveraged the 3.5B version of XCOMET [32][9], 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 a detected error spans field, which identifies potential errors and their positions within the sentence. To ensure translation quality, we manually reviewed and corrected all utterances that xCOMET had rated as "Weak" or "Moderate". Five annotators conducted the annotation following the same guidelines used during the initial verbalization phase (Appendix A.2). Their objective was to validate and, where necessary, revise the machine-translated verbalizations to ensure that the final benchmark constituted an high quality fully parallel dataset across the four languages. In particular, annotators were instructed to preserve the semantic content of the original data units while producing fluent and natural sentences that remained as faithful as possible to the input triple sets. Overall, we obtain a fully parallel dataset containing a total of 2,000 input (500 per language), some examples are presented in Figure 5 in Appendix. The final dataset included 1,376 machine-translated verbalizations, with an average xCOMET score of 0.81 suggesting an overall high quality. The complete results split per language for the remaining machine-translated verbalizations are reported in Table 1.

Language	XCOMET score ± SD	#items
Greek (EL)	0.794 ± 0.090	439
English (EN)	0.842 ± 0.069	327
French (FR)	0.782 ± 0.080	289
Italian (IT)	0.825 ± 0.079	321
Overall	**0.810 ± 0.084**	**1,376**

Table 1
Mean XCOMET quality estimation scores with standard deviations and numbers of evaluated machine translations by target language.

4. Methodology

To investigate whether multilingual RDF-to-Text systems exhibit gender or geographical bias, we evaluate six SoTA open-weights multilingual LLMs. Although our primary objective is to assess both gender and geographical biases, we also examine whether the geographical origin of the models influences their behavior. To this end, we select models developed in three different regions of the world, under the assumption that differences in their training data and development context may lead to distinct bias patterns. Specifically, our evaluation includes two models developed in America (Command-r-7B and Gemma4-E4B), two developed in Europe (Teuken-7B and EuroLLM-7B), and two developed in Asia (Qwen3.5-9B and GLM4-9B). The full model details and their Hugging Face links are provided in Table 4 in Appendix B.

[8] https://www.deepl.com/translator
[9] https://huggingface.co/Unbabel/XCOMET-XL

4.1. Experimental setup

Given a set of RDF-style triples *<subject, predicate, object>*, the system generates a single-paragraph natural language description in one of four target languages: English, French, Italian, or Greek. We evaluate the models under two prompting configurations: zero-shot and few-shot.

In the few-shot configuration, we need to minimize the influence of the examples on the models' outputs, since our dataset comprises researchers from diverse geographical regions. To this end, we select four demonstration examples: two from the most represented geographical category in the overall unbalances dataset (Southern Europe) and two from the least represented category (Middle Africa) of the overall dataset. Within each category, one example corresponds to a male researcher and the other to a female researcher, ensuring a balanced representation of gender. An example of the few shot examples are available in appendix B.2

Prompts are written in English, French, Italian, or Greek, depending on the target language verbalization. Full prompts are available in Appendix B.1.

4.2. Evaluation

To evaluate the results, we use multiple reference-based metrics commonly adopted in RDF-to-Text challenges [33, 34] to assess semantic similarity through embedding-based metrics and textual similarity through overlap-based metrics. In addition, we compute precision, recall, and F1-score to evaluate gender prediction accuracy. Specifically, we use BERTScore [35] to evaluate semantic similarity,[10] while textual similarity is assessed using BLEU [36], chrF++ [37], and ROUGE [38].[11] To compute precision, recall, and F1-score for gender prediction, we extract the gender from the generated verbalizations using the same criteria adopted for extracting the gold-reference gender from the OpenAIRE biographies (e.g., gender-marked pronouns or morphologically inflected words in languages with rich morphology, full details on gender extraction in table 5, Appendix B.3). Cases in which gender could not be determined from the generated verbalization (e.g., due to an incorrect verbalization or the absence of gender-marked pronouns or morphological cues) are labeled as "unknown" and excluded from the evaluation.

5. Results

The following section presents the results of the experiments. Sections 5.1 and 5.2 report the results obtained using embedding- and overlap-based evaluation metrics, with a particular focus on differences across languages and the gender of the input entities. Section 5.3 presents the results of the gender attribution analysis based on Precision, Recall, and F1-score.

5.1. Model Performance by Language

To provide an overview of system performance across languages, we computed the evaluation metrics introduced in Section 4.2 by comparing the generated verbalizations with the reference verbalizations in our datasets. Table 2 reports the results for each language. To assess the statistical significance of differences between models and prompting settings, we performed two-sided paired Wilcoxon signed-rank tests with Holm correction.

Overall, few-shot prompting consistently improves performance over zero-shot prompting, with statistically significant differences in most cases. This indicates that providing in-context examples generally

[10] We report BERTScore$_{F1}$ in the final results. We could not use the baseline-rescaled version because it is not available for Greek.

[11] We report ROUGE$_L$, which captures sentence-level structural similarity by identifying the longest common subsequence.

(a) Italian

Model	Prompt	BERTScore$_{F1}$	BLEU	chrF++	METEOR	ROUGE$_L$
Teuken-7B	ZS	0.749	0.127	40.342	0.318	0.286
	FS	0.759	0.249*	42.678	0.392*	0.395*
EuroLLM-9B	ZS	0.774	0.206	51.330	0.482	0.378
	FS	0.827*	0.335*	60.630*	0.594*	0.558*
Command-r-7B	ZS	0.808	0.255	52.030	0.480	0.420
	FS	**0.895***,†	**0.492***,†	**68.903***,†	**0.692***,†	**0.693***,†
Gemma4-E4B	ZS	0.831	0.311	58.662	0.576	0.478
	FS	0.842*	0.333*	60.449*	0.585*	0.585*
GLM4-9B	ZS	0.699	0.126	45.557	0.420	0.272
	FS	0.742*	0.224*	56.607*	0.583*	0.435*
Qwen3.5-9B	ZS	0.747	0.249	51.579	0.495	0.404
	FS	0.886*	0.447*	67.295*	0.671*	0.662*

(b) English

Model	Prompt	BERTScore$_{F1}$	BLEU	chrF++	METEOR	ROUGE$_L$
Teuken-7B	ZS	0.800	0.241	55.328	0.562	0.479
	FS	0.852*	0.407*	64.440*	0.672*	0.639*
EuroLLM-9B	ZS	0.789	0.211	56.073	0.575	0.425
	FS	0.832*	0.318*	60.929*	0.643*	0.581*
Command-r-7B	ZS	0.831	0.303	59.682	0.610	0.528
	FS	**0.912***,†	**0.533***,†	**74.722***,†	**0.805***,†	**0.792***,†
Gemma4-E4B	ZS	0.851	0.327	63.753	0.682	0.565
	FS	0.874*	0.389*	67.556*	0.718*	0.711*
GLM4-9B	ZS	0.738	0.123	48.419	0.502	0.294
	FS	0.739	0.133*	48.634	0.503	0.312*
Qwen3.5-9B	ZS	0.852	0.332	64.295	0.683	0.563
	FS	0.891*	0.429*	68.355*	0.740*	0.717*

(c) French

Model	Prompt	BERTScore$_{F1}$	BLEU	chrF++	METEOR	ROUGE$_L$
Teuken-7B	ZS	0.769	0.169	46.087	0.410	0.340
	FS	0.695*	0.113*	32.515*	0.265*	0.266*
EuroLLM-9B	ZS	0.759	0.172	49.410	0.473	0.326
	FS	0.797*	0.275*	55.073*	0.552*	0.471*
Command-r-7B	ZS	0.805	0.244	51.149	0.475	0.388
	FS	**0.899***,†	**0.479***,†	**67.815***,†	**0.707***,†	**0.666***,†
Gemma4-E4B	ZS	0.845	0.322	59.483	0.608	0.494
	FS	0.848*	0.350*	61.484*	0.626*	0.585*
GLM4-9B	ZS	0.712	0.117	45.017	0.419	0.260
	FS	0.717*	0.190*	52.783*	0.550*	0.382*
Qwen3.5-9B	ZS	0.771	0.247	51.423	0.498	0.396
	FS	0.881*	0.415*	65.399*	0.668*	0.621*

(d) Greek

Model	Prompt	BERTScore$_{F1}$	BLEU	chrF++	METEOR	ROUGE$_L$
Teuken-7B	ZS	0.708	0.063	32.235	0.275	0.369
	FS	0.721*	0.145*	35.252*	0.317*	0.445*
EuroLLM-9B	ZS	0.730	0.107	37.876	0.359	0.472
	FS	0.807*	0.360*	55.982*	0.582*	0.772*
Command-r-7B	ZS	0.738	0.156	35.957	0.370	0.498
	FS	0.832*	0.370*	53.170*	0.569*	0.651*
Gemma4-E4B	ZS	0.784	0.246	49.570	0.520	0.705
	FS	0.812*	0.369*	56.815*	0.626*	0.785*
GLM4-9B	ZS	0.693	0.086	34.821	0.318	0.422
	FS	0.791*	0.316*	60.458*	0.620*	0.628*
Qwen3.5-9B	ZS	0.748	0.093	34.047	0.333	0.369
	FS	**0.848***,†	**0.397***,†	**61.015***	**0.649***,†	**0.817***,†

Table 2
Automatic evaluation results for Italian, English, French, and Greek. ZS and FS denote zero-shot and few-shot prompting. Bold and underlined values indicate the best and second-best model–prompt configurations for each language and metric. * indicates a significant ZS–FS difference for the same model, while † indicates that the best configuration significantly outperforms the second-best (Holm-corrected two-sided paired Wilcoxon test, $p<0.05$).

leads to generated verbalizations that are more similar to the reference verbalizations according to the adopted automatic evaluation metrics.

Across the Italian, English, and French datasets, Command-r-7B achieves the best overall performance, significantly outperforming the second-best configuration in most metrics. Qwen3.5-9B consistently ranks second, showing competitive performance across all three languages.

The Greek results exhibit a different pattern. Qwen3.5-9B in the few-shot setting achieves the best overall performance, significantly outperforming the second-best configuration on all metrics except chrF++. Specifically, the second-best configuration varies across metrics: Command-r-7B ranks second in BERTScore$_{F1}$ and BLEU, Gemma4-E4B in METEOR and ROUGE$_L$, and GLM4-9B in chrF++.

5.2. Gender Gap in Verbalization Quality

We investigate whether the quality of the generated verbalizations differs according to the researchers' gender. Figure 1 reports the difference between the mean scores obtained for the Male and Female groups, expressed in score points on a common 0–100 scale. Positive values indicate higher performance for the Male group. Overall, most model–language–prompting configurations show positive gaps, suggesting a tendency toward slightly higher-quality verbalizations for male researchers. However, the magnitude and statistical significance of these differences vary substantially across languages, models, prompting strategies, and evaluation metrics.

The largest and most consistent disparities are observed for Greek. In particular, under few-shot prompting, Teuken-7B and EuroLLM-9B show marked Male–Female gaps across several metrics. Conversely, Command-r-7B and Gemma4-E4B display near-zero or negative gaps in the same setting. English shows smaller but relatively consistent male-favoring differences, whereas French appears more balanced and includes several configurations in which the Female group obtains slightly higher scores. Italian gener-

ally shows modest positive gaps, although larger descriptive differences are observed for Teuken-7B under few-shot prompting.

These results indicate that, in our experimental setting, few-shot prompting does not systematically mitigate gender-related performance disparities. Depending on the model and output language, it may reduce, preserve, or amplify the observed gaps. Moreover, since most differences are small and only a subset reaches statistical significance, these findings should be interpreted as evidence of heterogeneous gender-related differences in verbalization quality rather than as a uniform bias across all systems. Importantly, this analysis does not directly measure incorrect gender attribution. Gender realization is therefore examined separately through gender-specific precision, recall, and F1 scores in Section 5.3.

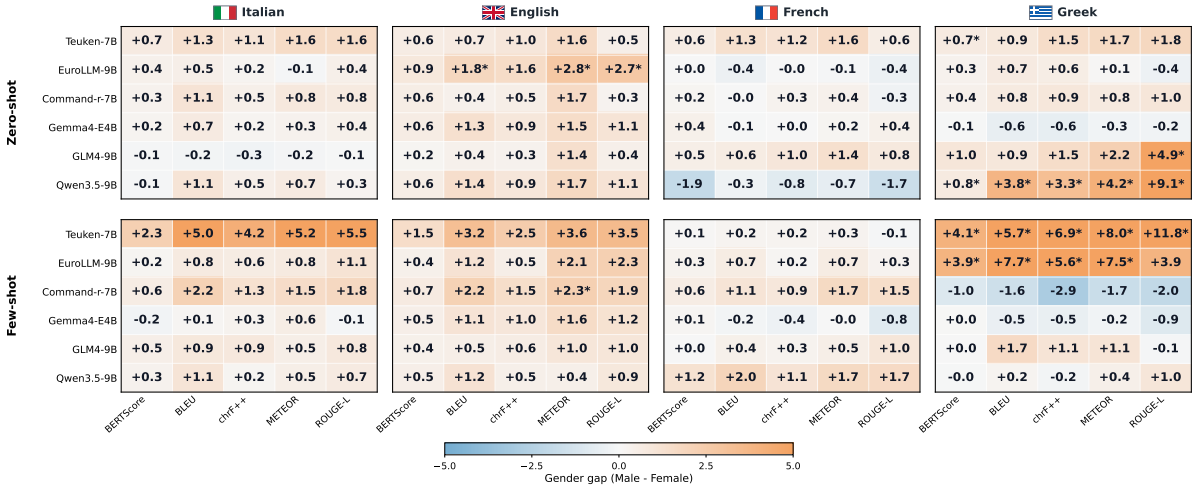


Figure 1: Gender gap in verbalization quality by prompting method, output language, model, and metric.Positive (orange) values indicate higher scores for the Male group, whereas negative (blue) values indicate higher scores for the Female group. An asterisk denotes significance after Holm correction across all evaluated model–metric comparisons within each language, based on two-sided Mann–Whitney U tests ($^*p < .05$).

5.3. Gender and Geographical Biases

To evaluate whether the generated verbalizations correctly identify the gender of the target researcher and whether significant differences exist between male and female researchers, we compute Precision, Recall, and F1-score for gender prediction. We then use two-sided Mann–Whitney U tests to assess the statistical significance of the observed differences. Considering all six models across the four languages, a total of 35,712 generations were produced for each prompting strategy (zero-shot and few-shot). The analysis was conducted on 30,739 generations, as 4,973 cases were excluded because the models failed to report any gender in the generated output (see Section 6 for details). Results are reported in Figure 2. Comparing languages, it is possible to notice that, overall, results in English are always better in respect to the other languages, whereas lower results are registered for Greek and Italian. Looking at the difference between male and female, it is interesting to notice that precision tends to be higher for female predictions, indicating that when a model predicts a name as female, it is generally correct. The performance drops when looking at recall for female values, which is consistently lower than the male values. This indicates that a high number of female inputs are misclassified as male.

To better investigate the male/female error differences we visualize the direction of gender attribution (Figure 3a) and the masculine overprediction gap (Figure 3b), which shows whether each model predicts the Male class more or less often than Male cases actually occur in the dataset. From Figure 3a, it clearly appears that the models tend to suffer from a male researcher bias, over-attributing the male gender in the verbalizations even when the researcher is female. This suggests that, in cases of uncertainty, the models tend to attribute male pronouns when referring to academic scholars. Figure 3b further confirms

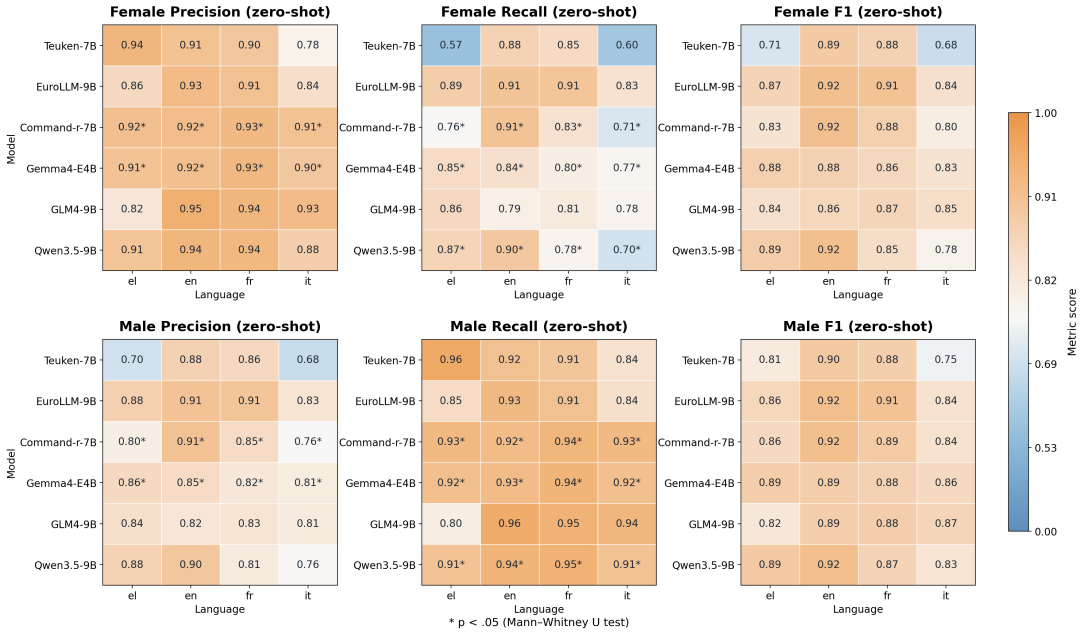


Figure 2: Results for the zero-shot setting of precision, recall, and F1 for gender extraction in Greek (el), English (en), French (fr), and Italian (it). Scores marked with * indicate statistically significant differences based on two-sided Mann–Whitney U tests ($*p < .05$).

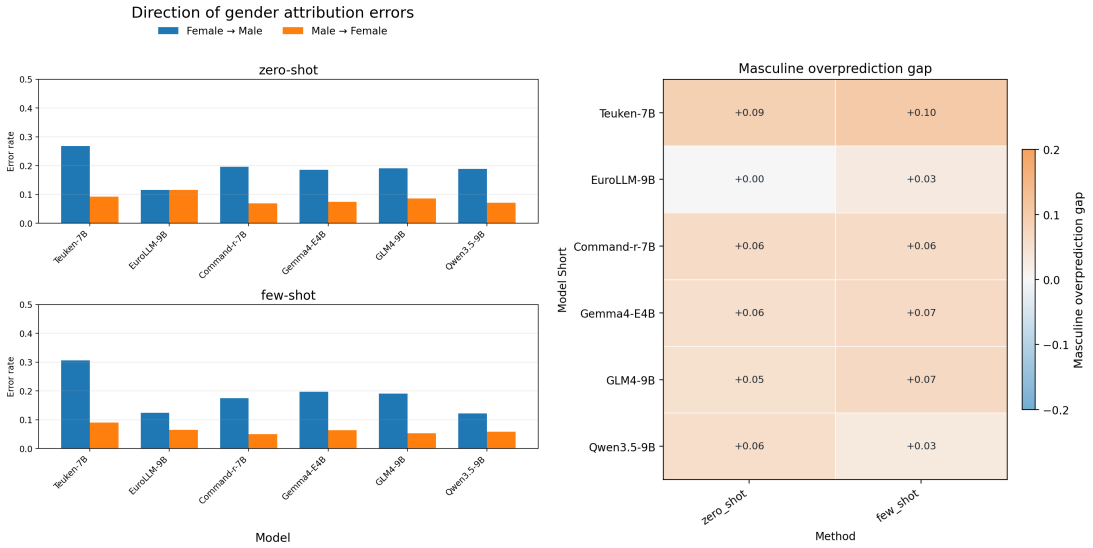
this trend: in all cases the values are positive, suggesting that all the models tend to overpredict the male gender. It is also worth noting that, for the EuroLLM model in the zero-shot setting, the value is +0.00, indicating that the proportion of male predictions matches the actual proportion of male subjects; the error rate increases, instead, in the few-shot setting.

Comparing researchers by the continent of their affiliated institution (Figure 4), we observe that gender prediction performance tends to be lower for researchers affiliated with institutions in Africa and Asia across all models, while it is generally higher for those affiliated with institutions in Europe and South America, the latter being predominantly composed of Spanish-speaking countries. Full table divided per gender are reported in Appendix C Tables 6a and 6b. No specific differences are found when comparing models developed in different areas, suggesting that there is an overall trend of misgendering researchers from non-Western countries. The two European models (Teuken-7B and EuroLLM-9B) exhibit the greatest variation across continents. In particular, Teuken-7B achieves the lowest accuracy in both the zero-shot and few-shot settings, whereas EuroLLM-9B attains some of the highest accuracies for researchers affiliated with institutions in Europe, South America, and North America

6. Error Analysis

To better understand the primary sources of error, we investigate which models exhibited the greatest difficulty in generating well-formed verbalizations, under which prompting strategy, and for researchers from which countries and gender.

First, we collected the sources of error by type; overall, across the three languages. On a total of 4,973 over the six models and the four languages detected errors we found 211 cases of empty prediction, 4,595 cases of no gender evidence, in which the gender is not extractable from the output, 17 cases of non-linguistic output (meaning that the output is written in scripts other than the ones analyzed), and 150 cases of too short output, in which the generated output contains only a few words. These statistics suggest that the generated outputs are predominantly well-formed, with only 378 malformed generations out of 4,973 represented cases. Therefore, the primary source of error appears to be the



(a) (b)

Figure 3: Figure (a) reports the directional gender attribution error rates by model and prompting strategy. The upper panel reports zero-shot results, the lower panel reports few-shot results. Female → Male denotes the proportion of Female instances incorrectly classified as Male, and Male → Female denotes the converse error. Figure (b) shows the Masculine overprediction gap (proportion predicted Male–proportion truly Male). Gap = 0.00: the proportion of Male predictions matches the actual proportion of Male subjects. Gap > 0: the model predicts Male too frequently. Gap < 0: the model predicts Male less frequently than expected.

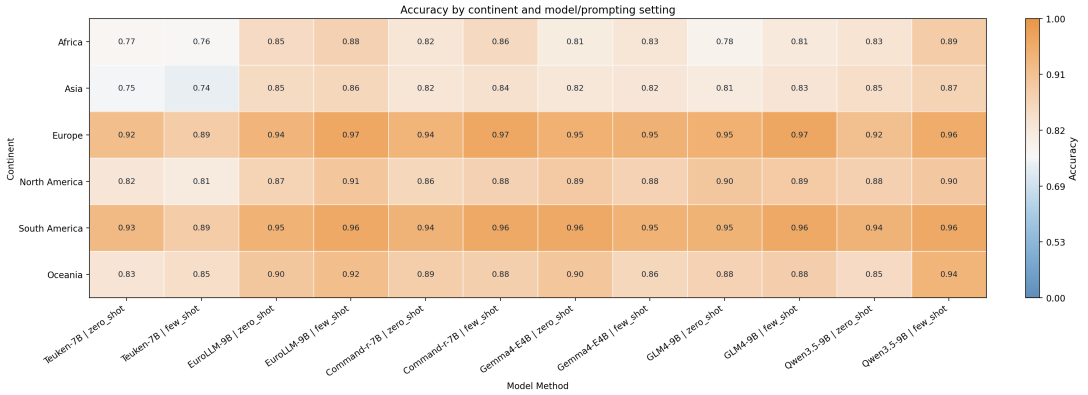


Figure 4: Comparison of gender-attribution accuracy across continents for each model under zero-shot and few-shot prompting. Accuracy is calculated as number of correct prediction / total number of cases.

models' inability to correctly identify the gender of the target researcher, rather than the generation of ill-formed verbalizations.

Overall, the model that shows the highest number of wrongly or not extracted genders is Teuken-7B (2583 error cases), generating in all the cases invalid output or simply rewriting the input triplets. Secondly, Command-r-7B presents the second-highest number of not assigned genders (922 cases), often outputting gender-unmarked pronouns such as "their", or using neutral word as "docente" in Italian, which can be both feminine and masculine. As in the sentence

> *Name Surname* is a researcher affiliated with Addis Ababa Science and Technology University. **Their** research focuses on topics such as clean water, diesel fuel, and potable water, with keywords including "Diesel Generator", "Potable Water", and "Water Pumping". **They** have published 10 papers, accumulating 54 citations and an H-index of 4.

It is also interesting to notice that the countries of affiliation that generate the major sources of errors

for GLM4-9B are Northern Africa, Southern Africa, and South-Eastern Asia, suggesting that the model struggles to correctly infer gender from proper names typical those regions.

Analyzing instead the cases in which models misgendered the researchers, the results show a similar trend: among the regions with a higher error rate are Eastern Asia, Southern Africa, and Northern Africa, and Teuken-7B is the model with the highest source of errors in both zero-shot and few-shot configurations. Teuken-7B tends to overgeneralize the male gender in its generations, even when dealing with proper names that are commonly used in Western countries and marked with "a", which is typically associated with female names. This is the case in the following examples (Italian), where the words "ricercatore" and "associato" are masculine terms, meaning "researcher" and "affiliated" in English:

> Mina Surname un **ricercatore** affiliato alla Tehran University of Medical Sciences, si occupa di ricerca sull'andrologia e la biologia...[12]

> Camilla Surname **associato** all'Università di Agder, svolge ricerca su argomenti come l'esaurimento e il business...[13]

Overall, it is possible to observe two opposite behaviors across models: some models, such as GLM4-9B, tend to remain gender-agnostic when the gender prediction is uncertain, producing answers that, while technically correct, do not correspond to the actual gender of the researcher; other models, instead, tend to overgeneralize the masculine gender, generating outputs that suffer from gender bias. This trend is even more visible when dealing with researchers from non-Western countries.

7. Conclusion and Future Works

In the present work, we conducted an in-depth investigation of gender and geographical biases on six different models, focusing on the RDF-to-Text task in the scholarly domain and considering researchers from three diverse geographical regions. Leveraging data extracted from well-established and reliable resources, such as OpenAIRE and SemOpenAlex, we assess whether SoTA LLMs could be reliable tools to help scholars in conducting open science initiatives, such as writing narrative curricula, or whether they could instead replicate biases or further marginalize less visible communities based on gender or geographical provenance.

Results show that, overall, all the tested models suffer from gender and geographical biases: there is a common trend to overgeneralize the masculine form over the feminine one, across all models and languages. In addition, models seem to struggle the most when dealing with African and Asian names, configuring a Western-centered bias. The qualitative analysis also reveals interesting insights, suggesting that different models exhibit different behaviors when dealing with entities for which they are uncertain about gender: GLM4-9B, for example, tends to produce outputs with gender-unmarked pronouns (e.g., "their"), while Teuken overgeneralizes the masculine form.

These results provide important insights into model behavior and can help scholars and stakeholders inform the choice of a particular tool. More generally, they can serve as a *caveat* when generating content that may be susceptible to different kinds of biases.

As future work, we plan to extend the size of the full dataset and to enrich the analysis with additional languages, especially outside the European area, in order to benchmark the models on several different languages with different gender markers. Furthermore, we will investigate bias mitigation techniques, including advanced approaches such as fine-tuning, to reduce gender bias in the models. We also plan

[12] English translation: Mina Surname, a researcher affiliated with Tehran University of Medical Sciences, conducts research in andrology and biology...

[13] English translation: Camilla Surname, an associate professor at the University of Agder, conducts research on topics such as burnout and business...

to enrich the evaluation strategy by adding proxies for the models' inherent biases, such as perplexity or conformal prediction.

Declaration on Generative AI

During the preparation of this work, the author(s) used AI tools in order to: Grammar and spelling check. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] B. Peng, Y. Zhu, Y. Liu, X. Bo, H. Shi, C. Hong, Y. Zhang, S. Tang, Graph retrieval-augmented generation: A survey, ACM Transactions on Information Systems 44 (2025) 1–52.

[2] P. Manghi, M. Artini, C. Atzori, M. Baglioni, A. Bardi, S. La Bruzzo, M. De Bonis, H. Dimitropoulos, I. Foufoulas, K. Iatropoulou, et al., Openaire: Advancing open science, in: Proceedings of the Nineteenth International Conference on Grey Literature, Rome, 2017, pp. 23–24.

[3] M. Färber, D. Lamprecht, J. Krause, L. Aung, P. Haase, Semopenalex: the scientific landscape in 26 billion rdf triples, in: International semantic web conference, Springer, 2023, pp. 94–112.

[4] L. Draetta, M. Oliverio, V. Ramón-Ferrer, P. F. Balestrucci, F. Corallo, C. Badenes-Olmedo, A. Mazzei, M. A. Stranisci, R. Damiano, Tailnlg: A multilingual benchmark addressing verbalization of long-tail entities, arXiv preprint arXiv:2603.27768 (2026).

[5] L. Draetta, M. A. Stranisci, F. Corallo, P. F. Balestrucci, M. Oliverio, R. Damiano, A. Mazzei, Beyond the metrics: an investigation into the reliability of evaluation metrics for domain specific graph-based question answering (2025).

[6] S. L. Blodgett, S. Barocas, H. Daumé Iii, H. Wallach, Language (technology) is power: A critical survey of "bias" in nlp, in: Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 5454–5476.

[7] D. Hovy, S. Prabhumoye, Five sources of bias in natural language processing, Language and linguistics compass 15 (2021) e12432.

[8] I. Staliūnaitė, J. Cheng, A. Vlachos, Uncertainty quantification for evaluating gender bias in machine translation, in: Findings of the Association for Computational Linguistics: EACL 2026, 2026, pp. 2204–2225.

[9] O. Menis Mastromichalakis, G. Filandrianos, M. Symeonaki, G. Stamou, Assumed identities: Quantifying gender bias in machine translation of gender-ambiguous occupational terms, in: C. Christodoulopoulos, T. Chakraborty, C. Rose, V. Peng (Eds.), Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2025, pp. 32233–32249. URL: https://aclanthology.org/2025.emnlp-main.1640/. doi:10.18653/v1/2025.emnlp-main.1640.

[10] S. M. Pawar, A. Arora, L.-A. Kaffee, I. Augenstein, Presumed cultural identity: How names shape LLM responses, in: C. Christodoulopoulos, T. Chakraborty, C. Rose, V. Peng (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2025, Association for Computational Linguistics, Suzhou, China, 2025, pp. 22147–22172. URL: https://aclanthology.org/2025.findings-emnlp.1207/. doi:10.18653/v1/2025.findings-emnlp.1207.

[11] H. Kotek, R. Dockum, D. Sun, Gender bias and stereotypes in large language models, in: Proceedings of the ACM collective intelligence conference, 2023, pp. 12–24.

[12] D. Duma, E. Klein, Generating natural language from linked data: Unsupervised template extraction, in: Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers, ASSOC COMPUTATIONAL LINGUISTICS-ACL, 2013, pp. 83–94.

[13] B. Distiawan, J. Qi, R. Zhang, W. Wang, 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), 2018, pp. 1627–1637.
[14] M. Oliverio, P. F. Balestrucci, A. Mazzei, V. Basile, 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, Association for Computational Linguistics, 2024, pp. 59–65.
[15] H. Gao, L. Wu, P. Hu, Z. Wei, F. Xu, B. Long, Rdf-to-text generation with reinforcement learning based graph-augmented structural neural encoders., CoRR (2021).
[16] S. Montella, B. Fabre, T. Urvoy, J. Heinecke, L. M. R. Barahona, 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+), 2020, pp. 89–99.
[17] F. Zhang, M. Zhang, S. Liu, Y. Sun, N. Duan, Enhancing rdf verbalization with descriptive and relational knowledge, ACM Transactions on Asian and Low-Resource Language Information Processing 22 (2023) 1–18.
[18] C. Gardent, A. Shimorina, S. Narayan, L. Perez-Beltrachini, The webnlg challenge: Generating text from rdf data, in: 10th International Conference on Natural Language Generation, ACL Anthology, 2017, pp. 124–133.
[19] T. Castro Ferreira, C. Gardent, N. Ilinykh, C. van der Lee, S. Mille, D. Moussallem, A. Shimorina (Eds.), Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+), Association for Computational Linguistics, Dublin, Ireland (Virtual), 2020. URL: https://aclanthology.org/2020.webnlg-1.0/.
[20] D. Torres-Salinas, E. Orduña-Malea, Á. Delgado-Vázquez, J. Gorraiz, W. Arroyo-Machado, Foundations of narrative bibliometrics, Journal of informetrics 18 (2024) 101546.
[21] D. Byrne, Could new'narrative'cvs transform research culture?, Nature (2023).
[22] A. D. SELBST, S. BAROCAS, Unfair artificial intelligence: How ftc intervention can overcome the limitations of discrimination law, University of Pennsylvania Law Review 171 (2023) 1023.
[23] J. Myung, N. Lee, Y. Zhou, J. Jin, R. A. Putri, D. Antypas, H. Borkakoty, E. Kim, C. Perez-Almendros, A. A. Ayele, et al., Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages, Advances in Neural Information Processing Systems 37 (2024) 78104–78146.
[24] Y. Y. Chiu, L. Jiang, B. Y. Lin, C. Y. Park, S. S. Li, S. Ravi, M. Bhatia, M. Antoniak, Y. Tsvetkov, V. Shwartz, et al., Culturalbench: a robust, diverse and challenging benchmark on measuring (the lack of) cultural knowledge of llms (2024).
[25] T. Alhanai, A. Kasumovic, M. M. Ghassemi, A. Zitzelberger, J. M. Lundin, G. Chabot-Couture, Bridging the gap: enhancing llm performance for low-resource african languages with new benchmarks, fine-tuning, and cultural adjustments, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, 2025, pp. 27802–27812.
[26] J. Etxaniz, G. Azkune, A. Soroa, O. L. de Lacalle, M. Artetxe, Bertaqa: How much do language models know about local culture?, Advances in Neural Information Processing Systems 37 (2024) 34077–34097.
[27] C. Gardent, A. Shimorina, S. Narayan, L. Perez-Beltrachini, Creating training corpora for nlg micro-planning, in: 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Association for Computational Linguistics (ACL), 2017, pp. 179–188.
[28] K. Aditya Hari, B. Singh, A. Sharma, V. Varma, WebNLG challenge 2023: Domain adaptive machine translation for low-resource multilingual RDF-to-text generation (WebNLG 2023), in: A. Gatt, C. Gardent, L. Cripwell, A. Belz, C. Borg, A. Erdem, E. Erdem (Eds.), Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023), Association for Computational Linguistics, Prague, Czech Republic, 2023, pp. 93–94. URL: https://aclanthology.org/2023.mmnlg-1.11/.
[29] M. Oliverio, P. F. Balestrucci, A. Mazzei, V. Basile, Webnlg-it: Construction of an aligned rdf-italian corpus through machine translation techniques, in: Findings of the Association for Computational Linguistics: ACL 2025, 2025, pp. 12073–12083.

[30] V. Ramón-Ferrer, C. Badenes-Olmedo, O. Corcho, Spanish triple-to-text benchmark on low-resource large language models, Semantic Web 17 (2026) 22104968261440701.
[31] D. Jurafsky, J. H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models, 3rd ed., 2025. URL: https://web.stanford.edu/~jurafsky/slp3/, online manuscript released August 24, 2025.
[32] N. M. Guerreiro, R. Rei, D. v. Stigt, L. Coheur, P. Colombo, A. F. T. Martins, xCOMET: Transparent machine translation evaluation through fine-grained error detection, Transactions of the Association for Computational Linguistics 12 (2024) 979–995. URL: https://aclanthology.org/2024.tacl-1.54/. doi:10.1162/tacl_a_00683.
[33] L. Cripwell, A. Belz, C. Gardent, A. Gatt, C. Borg, M. Borg, J. Judge, M. Lorandi, A. Nikiforovskaya, W. Soto-Martinez, C. Thomson, The 2023 WebNLG shared task on low resource languages. overview and evaluation results (WebNLG 2023), in: A. Gatt, C. Gardent, L. Cripwell, A. Belz, C. Borg, A. Erdem, E. Erdem (Eds.), Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023), Association for Computational Linguistics, Prague, Czech Republic, 2023, pp. 55–66. URL: https://aclanthology.org/2023.mmnlg-1.6/.
[34] S. Mille, J. Sedoc, Y. Liu, E. Clark, A. J. Axelsson, M. A. Clinciu, Y. Hou, S. Mahamood, I. N. Obonyo, L. Zhang, The 2024 GEM shared task on multilingual data-to-text generation and summarization: Overview and preliminary results, in: S. Mille, M.-A. Clinciu (Eds.), Proceedings of the 17th International Natural Language Generation Conference: Generation Challenges, Association for Computational Linguistics, Tokyo, Japan, 2024, pp. 17–38. URL: https://aclanthology.org/2024.inlg-genchal.2/. doi:10.18653/v1/2024.inlg-genchal.2.
[35] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, arXiv preprint arXiv:1904.09675 (2019).
[36] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
[37] M. Popović, chrf++: words helping character n-grams, in: Proceedings of the second conference on machine translation, 2017, pp. 612–618.
[38] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.

A. Dataset

Below, we report an example of a set of triples (i.e., a data unit) containing all the information extracted from the OpenAIRE and SemOpenAlex knowledge graphs.


```
[
  {
    "subject": "Name Surname",
    "predicate": "hasAffiliation",
    "object": "Universidade Federal de Pernambuco"
  },
  {
    "subject": "Name Surname",
    "predicate": "hasHIndex",
    "object": "1"
  },
  {
    "subject": "Name Surname",
    "predicate": "hasCitationCount",
    "object": "1"
  },
```

```
{
  "subject": "Name Surname",
  "predicate": "hasWorksCount",
  "object": "3"
},
{
  "subject": "Name Surname",
  "predicate": "hasKeyword",
  "object": "Dialogic"
},
{
  "subject": "Name Surname",
  "predicate": "hasKeyword",
  "object": "Queen"
},
{
  "subject": "Name Surname",
  "predicate": "hasKeyword",
  "object": "Temporalities"
},
{
  "subject": "Name Surname",
  "predicate": "conductsResearchOn",
  "object": "Art"
},
{
  "subject": "Name Surname",
  "predicate": "conductsResearchOn",
  "object": "Dialogic"
},
{
  "subject": "A Experiência Drag Queen como Transição na Vida Adulta",
  "predicate": "hasAuthor",
  "object": "Name Surname"
},
{
  "subject": "A Experiência Drag Queen como Transição na Vida Adulta",
  "predicate": "hasPublicationYear",
  "object": "2020"
},
{
  "subject": "A Experiência Drag Queen como Transição na Vida Adulta",
  "predicate": "hasCitationCount",
  "object": "1"
},
{
  "subject": "A drag queen e o gaymer como fissuras na cisheteronormatividade da cultura pop",
  "predicate": "hasAuthor",
  "object": "Name Surname"
},
{
  "subject": "A drag queen e o gaymer como fissuras na cisheteronormatividade da cultura pop",
  "predicate": "hasPublicationYear",
  "object": "2023"
},
{
  "subject": "A drag queen e o gaymer como fissuras na cisheteronormatividade da cultura pop",
  "predicate": "hasCitationCount",
```

```
    "object": "0"
  },
  {
    "subject": "Autobiographical Narrative as a Methodological Instrument of the Network
    of Meanings",
    "predicate": "hasAuthor",
    "object": "Name Surname"
  },
  {
    "subject": "Autobiographical Narrative as a Methodological Instrument of the Network
    of Meanings",
    "predicate": "hasPublicationYear",
    "object": "2023"
  },
  {
    "subject": "Autobiographical Narrative as a Methodological Instrument of the Network
    of Meanings",
    "predicate": "hasCitationCount",
    "object": "0"
  },
  {
    "subject": "Name Surname",
    "predicate": "sharesPaperWith",
    "object": "Renata Lira dos Santos Aléssio"
  },
  {
    "subject": "Name Surname",
    "predicate": "sharesPaperWith",
    "object": "Maria Isabel Pedrosa"
  },
  {
    "subject": "Name Surname",
    "predicate": "sharesPaperWith",
    "object": "Víctor Hugo"
  }
]
```


A.1. Dataset statistics

Table 3 reports the complete statistics of the dataset, including the minimum and maximum number of triples per data unit, the gender distribution of scholars, and their geographical provenance.

A.2. Verbalization guidelines

Here below we reported the Human verbalization guidelines that the annotators followed during both the manual verbalization and the correction of the input that had a low x-comet scores.

```
The aim of this task is to review the verbalizations generated by the LLMs and create
a reliable, human-validated ground truth that can serve as a reference for evaluating
the experimental results.

Your task is to carefully read the model outputs and revise them
according to the following guidelines:

- Write the final revised verbalization in the gold_verbalization   column.
- The final verbalization should sound as natural as possible, as if it were originally
  written by a human.
- All triples contained in the input data must be reflected in the verbalization.
```

Variable	Value
Triple Statistics	
Max triple number	25
Min triple number	13
Average triple number	21.83
Gender Distribution	
Male	249
Female	251
Entries per Country/Region	
Australia and New Zealand	26
Caribbean	26
Central America	26
Central Asia	26
Eastern Africa	24
Eastern Asia	26
Eastern Europe	28
Middle Africa	16
Northern Africa	28
Northern America	28
Northern Europe	28
South America	28
South-eastern Asia	28
Southern Africa	28
Southern Asia	26
Southern Europe	28
Western Africa	24
Western Asia	28
Western Europe	28
Total Entries	**500**

Table 3
Dataset Summary Statistics

```
 No information should be omitted.
- Since the focus of the project is on gender and geographical provenance, pay particular
  attention to these aspects and ensure that:
    - gender-specific information is correctly expressed (e,g,    pronouns and morphological
      markers)
    - Geographical information is correctly represented.

If you have any doubts, questions, or comments, please record them in the Note column.
Whenever possible, translate keywords into the target language, use your judgment
and preserve terms that are commonly used in English.
For example, "Sentiment Analysis" should remain in English.
```

A.3. Translation

In figure 5 are reported one input example translated in the four languages.

B. Experimental Setup

All experiments were conducted on a QuantaGrid S74G-2U server based on the `aarch64` architecture, featuring a Grace A02 CPU @ 3.4GHz with 72 cores, 480GB of RAM, one NVIDIA H100 GPU, and a 100Gbps Ethernet connection. All local decoder models listed in Table 4 were loaded and run locally using the Hugging Face `Transformers` Python package. Generation was performed ...

Multilingual Verbalization Example

EN:
Luis Fernando León-Fernandez is a researcher affiliated with the Flemish Institute for Technological Research. His research focuses on topics such as Agriculture and Atmosphere (unit), using keywords like Interface and Watershed Management. He has published 39 works, accumulating 461 citations and an H-Index of 14. Among his most cited contributions are "Towards coupling a 3D hydrodynamic lake model with the Canadian Regional Climate Model: Simulation on Great Slave Lake" (2006, 49 citations), "Integration of a nonpoint source pollution model with a decision support system" (2000, 40 citations) and "Watershed management modelling in Malawi: application and technology transfer" (2003, 19 ciations), He frequently collaborates with David Lam, D SWAYNE and J. Imberger

IT:
Luis Fernando León-Fernandez è un ricercatore affiliato all'Istituto fiammingo per la ricerca tecnologica. La sua attività di ricerca si concentra su temi quali Agricoltura e Atmosfera (unità), utilizzando parole chiave quali Interfaccia e Gestione dei bacini idrografici. Ha pubblicato 39 lavori, totalizzando 461 citazioni e un indice H pari a 14. Tra i suoi contributi più citati figurano «Towards coupling a 3D hydrodynamic lake model with the Canadian Regional Climate Model: Simulation on Great Slave Lake» (2006, 49 citazioni), «Integration of a nonpoint source pollution model with a decision support system» (2000, 40 citazioni) e "Modellizzazione della gestione dei bacini idrografici in Malawi: applicazione e trasferimento tecnologico" (2003, 19 citazioni). Collabora frequentemente con David Lam, D. SWAYNE e J. Imberger.

FR:
Luis Fernando León-Fernandez est chercheur affilié à l'Institut flamand de recherche technologique. Ses travaux portent notamment sur des thèmes tels que l'agriculture et l'atmosphère (unité), autour de mots-clés comme « interface » et « gestion des bassins versants ». Il a publié 39 articles, totalisant 461 citations et affichant un indice H de 14. Parmi ses contributions les plus citées figurent « Towards coupling a 3D hydrodynamic lake model with the Canadian Regional Climate Model: Simulation on Great Slave Lake » (2006, 49 citations), « Integration of a nonpoint source pollution model with a decision support system » (2000, 40 citations) et « Modélisation de la gestion des bassins versants au Malawi : application et transfert de technologie » (2003, 19 citations). Il collabore fréquemment avec David Lam, D. SWAYNE et J. Imberger.

EL:
Ο Luis Fernando León-Fernandez είναι ερευνητής που συνεργάζεται με το Φλαμανδικό Ινστιτούτο Τεχνολογικής Έρευνας. Η έρευνά του επικεντρώνεται σε θέματα όπως η Γεωργία και η Ατμόσφαιρα (μονάδα), με λέξεις-κλειδιά όπως «Διεπαφή» και «Διαχείριση Λεκανών Απορροής». Έχει δημοσιεύσει 39 εργασίες, οι οποίες έχουν συγκεντρώσει 461 αναφορές και έχουν δείκτη Η ίσο με 14. Μεταξύ των πιο συχνά αναφερόμενων εργασιών του συγκαταλέγονται τα εξής: «Towards coupling a 3D hydrodynamic lake model with the Canadian Regional Climate Model: Simulation on Great Slave Lake» (2006, 49 αναφορές), «Integration of a nonpoint source pollution model with a decision support system» (2000, 40 αναφορές) και «Μοντελοποίηση της διαχείρισης λεκανών απορροής στο Μαλάουι: εφαρμογή και μεταφορά τεχνολογίας» (2003, 19 αναφορές). Συνεργάζεται συχνά με τους David Lam, D. SWAYNE και J. Imberger.

Figure 5: Example of multilingual verbalization generated from the same data units in English, Italian, French, and Greek.

Model	Referred to as
Teuken-7B-instruct-v0.6	Teuken-7B
EuroLLM-9B-Instruct-2512	EuroLLM-9B
c4ai-Command-r-7B-12-2024	Command-r-7B
gemma-4-E4B-it	Gemma4-E4B
glm-4-9b-chat-hf	GLM4-9B
Qwen3.5-9B	Qwen3.5-9B

Table 4
Decoder models used in this work, together with the naming convention adopted throughout the paper.

B.1. Prompt

English Prompt

```
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.
- Verbalize all input triples, covering every piece of information in a fluent paragraph.
- Continue after the provided prefix with only the final verbalization paragraph.
- Do not add labels, bullet points, explanations, or alternative outputs.
- Do not write reasoning, chain-of-thought, or <think> blocks.
- Do not copy the triples and do not use the [subject, predicate, object] format in the answer.
INPUT TRIPLES: {triples}

Italian Prompt

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.
ISTRUZIONI:
- Genera il testo esclusivamente dalle triple in input. - Verbalizza tutte le triple in input, includendo ogni informazione in un paragrafo fluente.
- Continua dopo il prefisso fornito con il solo paragrafo finale di verbalizzazione.
- Non aggiungere etichette, elenchi puntati, spiegazioni o output alternativi.
- Non scrivere ragionamenti, chain-of-thought o blocchi <think>.
- Non copiare le triple e non usare il formato [soggetto, predicato, oggetto] nella risposta.
TRIPLE IN INPUT: {triples}

French Prompt

En français, les données structurées sont couramment représentées sous forme de triplets, au format [sujet, prédicat, objet]. À partir de ces triplets, générez un texte en un seul paragraphe composé de phrases complètes, grammaticalement correctes et naturelles.
INSTRUCTIONS:
- Générez le texte uniquement à partir des triplets en entrée.
- Verbalisez tous les triplets en entrée, en couvrant chaque information dans un paragraphe fluide.
- Continuez après le préfixe fourni avec uniquement le paragraphe final de verbalisation.
- N'ajoutez pas d'étiquette, de liste à puces, d'explication ni de sortie alternative.
- N'écrivez pas de raisonnement, de chaîne de pensée ni de blocs <think>.
- Ne copiez pas les triplets et n'utilisez pas le format [sujet, prédicat, objet] dans la réponse.
TRIPLETS EN ENTRÉE : {triples}

Greek Prompt

Στα ελληνικά, τα δομημένα δεδομένα αναπαρίστανται συνήθως ως τριάδες, με τη μορφή [υποκείμενο, κατηγόρημα, αντικείμενο]. Με βάση αυτές τις τριάδες, δημιουργήστε ένα κείμενο μίας παραγράφου που αποτελείται από πλήρεις, γραμματικά σωστές και φυσικές προτάσεις.
ΟΔΗΓΙΕΣ:
- Δημιουργήστε το κείμενο αποκλειστικά από τις τριάδες εισόδου.
- Λεκτικοποιήστε όλες τις τριάδες εισόδου, καλύπτοντας κάθε πληροφορία σε μια ρέουσα παράγραφο.
- Συνεχίστε μετά το δοσμένο πρόθεμα μόνο με την τελική παράγραφο λεκτικοποίη-

σης.
- Μην προσθέτετε ετικέτες, κουκκίδες, εξηγήσεις ή εναλλακτικές απαντήσεις.
- Μην γράφετε συλλογισμό, chain-of-thought ή μπλοκ <think>.
- Μην αντιγράφετε τις τριάδες και μην χρησιμοποιείτε τη μορφή [υποκείμενο, κατηγόρημα, αντικείμενο] στην απάντηση.
ΤΡΙΑΔΕΣ ΕΙΣΟΔΟΥ: {triples}

B.2. Few-shot Examples

```
{"subject": "Eshan Karunarathne", "predicate": "hasAffiliation", "object": "University of Melbourne
    "}
{"subject": "Eshan Karunarathne", "predicate": "hasHIndex", "object": "7"}
{"subject": "Eshan Karunarathne", "predicate": "hasCitationCount", "object": "166"}
{"subject": "Eshan Karunarathne", "predicate": "hasWorksCount", "object": "14"}
{"subject": "Eshan Karunarathne", "predicate": "hasKeyword", "object": "Minification"}
{"subject": "Eshan Karunarathne", "predicate": "hasKeyword", "object": "Overvoltage"}
{"subject": "Eshan Karunarathne", "predicate": "hasKeyword", "object": "Premature Convergence"}
{"subject": "Eshan Karunarathne", "predicate": "conductsResearchOn", "object": "AC power"}
{"subject": "Eshan Karunarathne", "predicate": "conductsResearchOn", "object": "Algorithm"}
{"subject": "Optimal Placement and Sizing of DGs in Distribution Networks Using MLPSO Algorithm", "
    predicate": "hasAuthor", "object": "Eshan Karunarathne"}
{"subject": "Optimal Placement and Sizing of DGs in Distribution Networks Using MLPSO Algorithm", "
    predicate": "hasPublicationYear", "object": "2020"}
{"subject": "Optimal Placement and Sizing of DGs in Distribution Networks Using MLPSO Algorithm", "
    predicate": "hasCitationCount", "object": "53"}
{"subject": "Mitigation of overvoltage due to high penetration of solar photovoltaics using smart
    inverters volt/var control", "predicate": "hasAuthor", "object": "Eshan Karunarathne"}
{"subject": "Mitigation of overvoltage due to high penetration of solar photovoltaics using smart
    inverters volt/var control", "predicate": "hasPublicationYear", "object": "2020"}
{"subject": "Mitigation of overvoltage due to high penetration of solar photovoltaics using smart
    inverters volt/var control", "predicate": "hasCitationCount", "object": "19"}
{"subject": "Comprehensive learning particle swarm optimization for sizing and placement of
    distributed generation for network loss reduction", "predicate": "hasAuthor", "object": "Eshan
    Karunarathne"}
{"subject": "Comprehensive learning particle swarm optimization for sizing and placement of
    distributed generation for network loss reduction", "predicate": "hasPublicationYear", "object
    ": "2020"}
{"subject": "Comprehensive learning particle swarm optimization for sizing and placement of
    distributed generation for network loss reduction", "predicate": "hasCitationCount", "object":
    "11"}
{"subject": "Eshan Karunarathne", "predicate": "sharesPaperWith", "object": "Janaka Ekanayake"}
{"subject": "Eshan Karunarathne", "predicate": "sharesPaperWith", "object": "Jagadeesh Pasupuleti"}
{"subject": "Eshan Karunarathne", "predicate": "sharesPaperWith", "object": "Dilini Almeida"}
```

IT: Eshan Karunarathne è un ricercatore affiliato all'University of Melbourne. La sua ricerca si concentra su AC power e algoritmi, con keyword tra cui minification, overvoltage e premature convergence. Ha pubblicato 14 lavori, accumulando 166 citazioni e un H-Index pari a 7. Tra i suoi contributi più citati figurano "Optimal Placement and Sizing of DGs in Distribution Networks Using MLPSO Algorithm" (2020, 53 citazioni), "Mitigation of overvoltage due to high penetration of solar photovoltaics using smart inverters volt/var control" (2020, 19 citazioni) e "Comprehensive learning particle swarm optimization for sizing and placement of distributed generation for network loss reduction" (2020, 11 citazioni). Collabora frequentemente con Janaka Ekanayake, Jagadeesh Pasupuleti e Dilini Almeida.

FR: Eshan Karunarathne est un chercheur affilié à l'University of Melbourne. Il a publié 14 travaux, totalisant 166 citations et un H-Index de 7. Ses recherches portent sur la minimisation, l'algorithme, le réseau électrique AC et la conversion de tension prématurée. Parmi ses contributions les plus citées figurent "Optimal Placement and Sizing of DGs in Distribution Networks Using MLPSO Algorithm" (2020, 53 citations), "Mitigation of overvoltage due to high penetration of solar photovoltaics using smart inverters volt/var control" (2020, 19 citations) et "Comprehensive learning particle swarm optimization for sizing and placement of distributed generation for

Language	Category	Weight	Example cues
English	Pronouns	3.0	*he, him, his*; *she, her, hers*
	Gendered profession nouns	3.0	*male researcher*, *female scientist*
Italian	Pronouns	3.0	*lui, egli*; *lei, ella*
	Article + profession title	2.5	*il ricercatore*, *la ricercatrice*
	Profession title	2.0	*autore*, *autrice*, *professore*, *professoressa*
	Gender morphology	1.0	*affiliato/affiliata*, *noto/nota*
French	Pronouns	3.0	*il, lui*; *elle*
	Article + profession title	2.5	*le chercheur*, *la chercheuse*
	Profession title	2.0	*chercheur*, *chercheuse*, *autrice*
	Plural pronouns	2.0	*ils*, *eux*
	Gender morphology	1.0	*affilié/affiliée*, *connu/connue*
Greek	Pronouns	3.0	*αυτός*; *αυτή*
	Possessives	2.5	*του*, *της*
	Article + profession title	2.5	*ο ερευνητής*, *η ερευνήτρια*
	Profession title	2.0	*καθηγητής*, *καθηγήτρια*
	Subject + possessive	2.5	*η έρευνα του*, *η έρευνα της*
	Gender morphology	1.0	*γνωστός/γνωστή*

Table 5
Rule categories used for multilingual gender inference. Each rule consists of a regular expression associated with a confidence weight.

network loss reduction" (2020, 11 citations). Il collabore fréquemment avec Janaka Ekanayake, Jagadeesh Pasupuleti et Dilini Almeida.

EN: Eshan Karunarathne is a researcher affiliated with the University of Melbourne. He has an H-Index of 7 and a citation count of 166, based on his 14 published works. His research interests include minification, overvoltage, and premature convergence, particularly in the context of AC power and algorithms. Some of his notable contributions include "Optimal Placement and Sizing of DGs in Distribution Networks Using MLPSO Algorithm" (2020, 53 citations), "Mitigation of overvoltage due to high penetration of solar photovoltaics using smart inverters volt/var control" (2020, 19 citations), and "Comprehensive learning particle swarm optimization for sizing and placement of distributed generation for network loss reduction" (2020, 11 citations). He often collaborates with Janaka Ekanayake, Jagadeesh Pasupuleti, and Dilini Almeida.

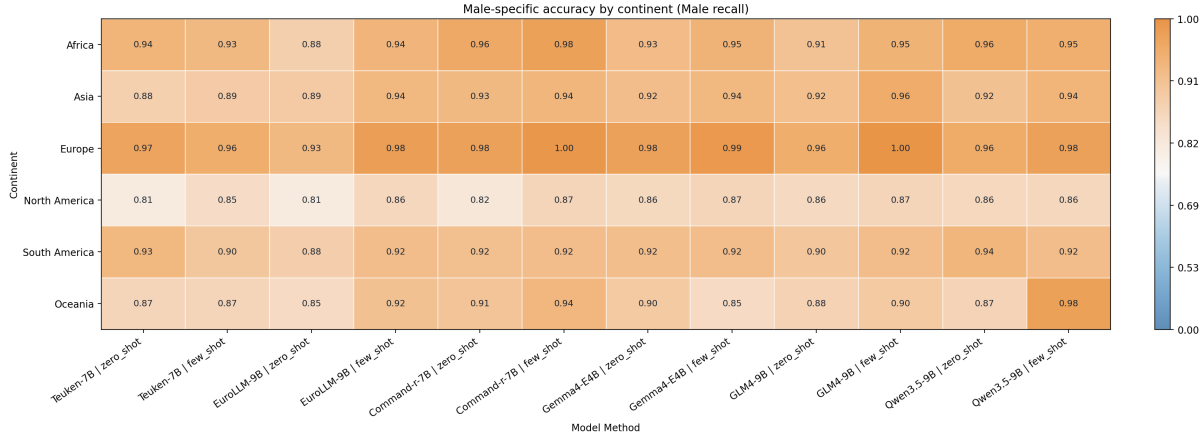
EL: Eshan Karunarathne είναι ένας ερευνητής που ανήκει στο University of Melbourne. Έχει H-Index 7, 166 αναφορές και έχει δημοσιεύσει 14 εργασίες. Μεταξύ των λέξεων-κλειδιών του είναι το minification, overvoltage και premature convergence. Έχει συγχρόνως επικεντρώνει τη θέματα AC power και algorithm. Μεταξύ των πιο δημοφιλημένων εργασιών του βρίσκονται ``Optimal Placement and Sizing of DGs in Distribution Networks Using MLPSO Algorithm'' (2020, 53 αναφορές), ``Mitigation of overvoltage due to high penetration of solar photovoltaics using smart inverters volt/var control'' (2020, 19 αναφορές) και ``Comprehensive learning particle swarm optimization for sizing and placement of distributed generation for network loss reduction'' (2020, 11 αναφορές). Συνεργάζεται συχνά με Janaka Ekanayake, Jagadeesh Pasupuleti και Dilini Almeida.

B.3. Evaluation

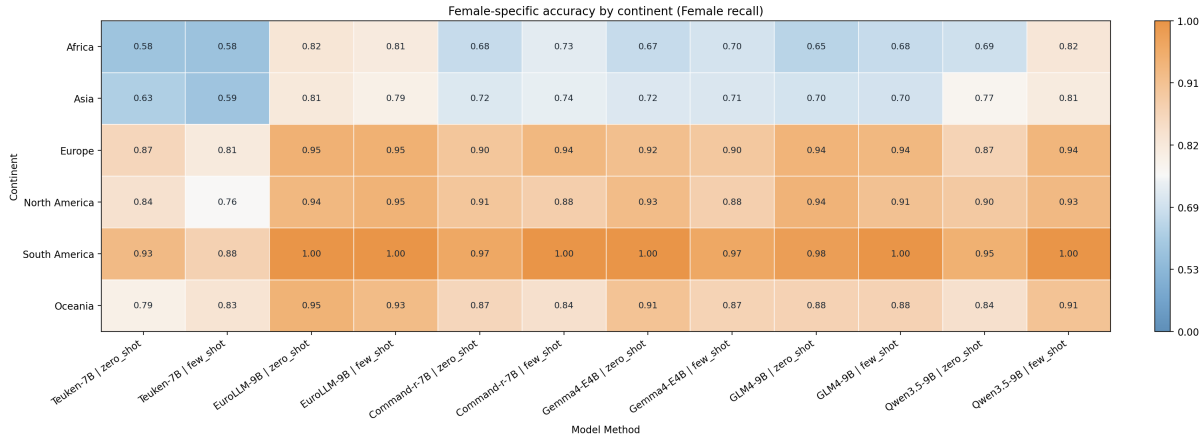
To extract gender information from the verbalizations and compute precision, recall, and F1-score, we adopted a language-specific string-matching approach based on manually defined lexical patterns. For each language, we searched for gender-specific pronouns and lexical markers. Table 5 report the patterns.

C. Results

Tables 6a and 6b show the accuracy of gender prediction over different geographical area divvied per gender (male and female).



(a)



(b)

Figure 6: Comparison of gender-attribution accuracy across continents for male (a) and female (b) inputs under zero-shot and few-shot prompting.