MORPHOGEN: A Multilingual Benchmark for Evaluating Gender-Aware Morphological Generation

Aditya Aggarwal[1*] **Mehul Agarwal**[1*] **Arnav Goel**[2†] **Medha Hira**[2†] **Anubha Gupta**[1‡]

[1]SBILab, Indraprastha Institute of Information Technology Delhi
[2]Carnegie Mellon University

{aditya22028, mehul22294}@iiitd.ac.in
{arnavgoe, mhira}@cs.cmu.edu
anubha@iiitd.ac.in

Abstract

While multilingual large language models (LLMs) perform well on high-level tasks like translation and question answering, their ability to handle grammatical gender and morphological agreement remains underexplored. In morphologically rich languages, gender influences verb conjugation, pronouns, and even first-person constructions with explicit and implicit mentions to gender. We introduce MORPHOGEN a morphologically grounded large-scale benchmark dataset for evaluating gender-aware generation in three typologically diverse grammatically gendered languages i.e. French, Arabic and Hindi. The core task, GENFORM, requires models to rewrite a first-person sentence in the opposite gender while preserving its meaning and structure. We construct a high-quality synthetic dataset spanning French, Arabic, and Hindi, and benchmark 15 popular multilingual LLMs (2B–70B) on their ability to perform this transformation. Our results reveal gaps and interesting insights into the handling of morphological gender in current models. MORPHOGEN offers a focused diagnostic lens for gender-aware language modeling and lays the groundwork for future research on inclusive and morphology-sensitive multilingual LLMs.

1 Introduction

Multilingual large language models (LLMs) achieve strong performance across tasks such as summarization, translation, and question answering [1, 2, 3, 4, 5]. Standard benchmarks like XTREME [6], Global-MMLU [7], and BenchMAX [8] have enabled broad evaluation, but face limitations including translation errors, contamination, and an emphasis on high-level semantic tasks. This makes it difficult to isolate fine-grained weaknesses in morphologically rich or gendered languages [9].

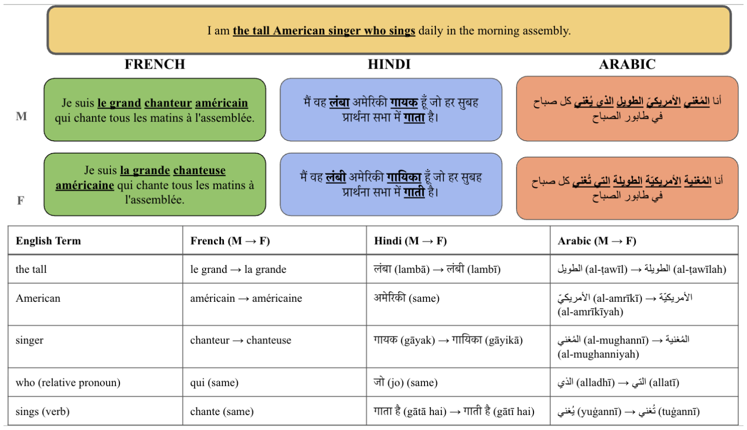
As LLMs are deployed in diverse linguistic settings, it is essential to evaluate their ability to apply grammatical rules consistently. This is especially critical for languages such as French, Arabic, and Hindi, where gender affects verbs, pronouns, and adjectives [10, 11]. Accurate modeling of gender morphology is vital both for inclusive applications like conversational agents and translation, and for probing bias in gendered language structures [12, 13].

However, no existing benchmark directly evaluates whether LLMs can generate coherent, grammatical sentences conditioned on gender. To address this, we introduce MORPHOGEN, a benchmark for gender-

[*] Equal contribution.

[†] Equal advising

[‡] Corresponding author.


I am the tall American singer who sings daily in the morning assembly.
FRENCH
HINDI
ARABIC
M
Je suis le grand chanteur américain qui chante tous les matins à l'assemblée.
मैं वह लंबा अमेरिकि गायक हूँ जो हर सुबह प्रार्थना सभा में गाता है।
أنا المُغني الأمريكيّ الطويل الذي يُغني كل صباح في طابور الصباح
F
Je suis la grande chanteuse américaine qui chante tous les matins à l'assemblée.
मैं वह लंबी अमेरिकि गायिका हूँ जो हर सुबह प्रार्थना सभा में गाती है।
أنا المُغنية الأمريكيّة الطويلة التي تُغني كل صباح في طابور الصباح

English Term	French (M → F)	Hindi (M → F)	Arabic (M → F)
the tall	le grand → la grande	लंबा (lambā) → लंबी (lambī)	الطويل (al-ṭawīl) → الطويلة (al-ṭawīlah)
American	américain → américaine	अमेरिकि (same)	الأمريكيّ (al-amrīkī) → الأمريكيّة (al-amrīkīyah)
singer	chanteur → chanteuse	गायक (gāyak) → गायिका (gāyikā)	المُغني (al-mughannī) → المُغنية (al-mughanniyah)
who (relative pronoun)	qui (same)	जो (jo) (same)	الذي (alladhī) → التي (allatī)
sings (verb)	chante (same)	गाता है (gātā hai) → गाती है (gātī hai)	يُغني (yuġannī) → تُغني (tuġannī)

Figure 1: Example illustrating how gender-based morphology differs across the three languages

conditioned morphological reasoning in French, Arabic, and Hindi. We define the GENFORM task: given a sentence and speaker gender, the model must rewrite it in the opposite gender while preserving correctness and meaning. Constructed from systematic gender-marking rules, the dataset poses compositional challenges beyond surface transformations. We evaluate 15 open- and closed-source multilingual LLMs ranging from 4B to 70B parameters.

Our contributions are: **(1)** A new benchmark covering three typologically diverse languages, with parallel English data to support translation and bias analysis (Section 2); **(2)** Novel evaluation metrics for assessing gender transformations, also applicable to translation and bias detection (Section 3.2); and **(3)** A comprehensive evaluation of multilingual LLMs on the GENFORM task, providing insights into their capacity to model gendered morphology.

2 Dataset

We introduce the MORPHOGEN dataset, a benchmark designed to evaluate the ability of multilingual LLMs to handle morphologically grounded gender transformations. It spans three typologically diverse languages—French, Arabic, and Hindi—that differ significantly in how gender is marked and propagated, thereby providing a robust testbed for probing cross-lingual generalization. Each sentence is paired with its gender counterfactual (masculine ⇌ feminine), enabling precise evaluation of whether models can transform gendered terms without altering unrelated content. A detailed description of language-specific gender morphology and dataset construction is provided in Appendix B.

2.1 Dataset Description and Design

Statistics	Arabic	French	Hindi
Unique Sentences	2,719	9,999	7,610
Number of Rules	14	12	13
Avg. Gender Terms*	2.02	1.78	1.43
Max. Gender Terms*	7	7	7
Avg. Word Count*	12.34	26.76	15.46
Max. Word Count*	38	67	87

Table 1: Statistics for MORPHOGEN (*per sentence)

MORPHOGEN covers 9,999 French, 2,719 Arabic, and 7,610 Hindi sentences (Table 1). Every instance consists of (i) a masculine sentence, (ii) its feminine counterpart, and (iii) an English parallel. **Gendered terms**—the tokens differing between the two gendered variants—serve as ground truth units for evaluation. Figure 4 shows their per-sentence distribution, with some sentences containing up to seven, underscoring the morphological complexity.

The dataset spans a diverse range of grammatical contexts where gender marking arises: *verbs and tense-specific agreement*, *adjectives and occupations*, *pronouns and possessives*, and *clause-level structures* such as passive voice and object-fronting. We additionally include **multi-entity sentences**, where only the speaker's gender governs agreement, to test robustness against *gender interference*. Together, these design choices ensure that MORPHOGEN systematically captures the morphosyntactic phenomena central to gender realization across languages. Further details and rule inventories are in Appendix B.

2.2 Task Formulation

The benchmark defines a controlled sentence rewriting task: given a first-person input sentence, the model must output the same sentence in the opposite gender. Successful completion requires applying language-specific morphological rules (e.g., suffixation in Arabic, agreement in French, natural-gender suffixes in Hindi) while preserving meaning, fluency, and syntactic structure. Figure 2 illustrates the general morphological processes underlying gender transformations across the three languages. This formulation makes `MORPHOGEN` a fine-grained test of LLMs' sensitivity to grammatical gender, disentangling correct rule application from unintended edits, and providing systematic coverage of both simple and compositionally complex cases.

3 Experimental Setup

3.1 Models Benchmarked

We evaluate 15 multilingual LLMs spanning major model families (LLAMA, Qwen, Gemma, and Phi) and parameter scales from 2B to 70B. This covers both lightweight models suited for deployment and large-capacity models expected to generalize well across languages. Full model specifications are provided in Appendix C.1.

3.2 Evaluation Metrics

To assess gender-conditioned transformations in `MORPHOGEN`, we design three complementary metrics:

- **Sentence-Level Gender Accuracy (SGA):** proportion of gendered terms correctly modified per sentence, averaged across the corpus.
- **Gender IoU (GIoU):** stricter measure penalizing both missed and spurious gender changes, inspired by Intersection-over-Union.
- **Corpus-Level Gender Accuracy (CGA):** overall ratio of correct gendered terms across the test set.

These metrics capture accuracy, precision–recall tradeoffs, and corpus-level quality. Full definitions and equations are deferred to Appendix C.2.

4 Results and Discussion

We evaluated 15 popular open-source and close-source multilingual LLMs on the `MORPHOGEN` benchmark across French, Arabic and Hindi, using the metrics proposed in Section 3.2. Results are presented language-wise as follows: Hindi in Table 8, French in Table 9 and Arabic in Table 10. We provide a detailed analysis on the variation in performance of different models (sizes and family) on `MORPHOGEN` and the corresponding inferences on gender bias in these models.

4.1 Smaller LMs can't handle Complex Morphology

Larger models consistently outperformed smaller ones across all languages, particularly in Arabic, where increased parameter size mitigated morphological complexity. For example, `Gemma3-27B` (27B parameters) achieved a CGA of 74.74% in Arabic, markedly outperforming `Gemma2-2B` at 14.10%. In Hindi, smaller models remained viable due to simpler rules, with `LLAMA-3.1-8B` scoring a CGA of 89.21%, compared to `LLAMA-3.3-70B` at 91.40%. French's larger dataset challenged resource-constrained models, amplifying errors, as `Gemma2-2B` recorded a CGA of 37.54%, while `Phi4-14B` reached 87.70%. This suggests that parameter size is critical for handling complex morphology but less impactful in simpler linguistic contexts like Hindi.

4.2 Masculine Bias in French and Arabic

Gender bias varied notably across languages, as seen in the $\triangle$SGA scores. In Hindi, bias was generally low but occasionally skewed toward feminine forms, with models like `Gemma3-4B` showing an $\triangle$SGA of `-14.32%`, often preferring feminine outputs even when the target gender was male. In French, a stronger masculine bias was observed, particularly in larger models such as `LLaMA3-70B`, which exhibited an $\triangle$SGA of `15.15%` due to consistent defaulting to masculine forms. Arabic showed persistent masculine bias, especially in plural constructions Figure 9 of appendix, with Qwen3-32B recording an $\triangle$SGA of 11.94%, frequently generating masculine outputs even in all-female contexts. These trends highlight the influence of gender bias of the training data used in these LLM's and underscore the need for targeted debiasing in morphologically rich languages.

4.3 Significant Variance in Model Families

Architectural differences influenced performance quality. Gemma models excelled in gender fairness, particularly in Arabic, maintaining balance in complex contexts. LLAMA models showed consistency in Hindi and French but struggled with bias in Arabic. Qwen models frequently exhibited masculine bias across languages, suggesting weaker gender handling. Phi models achieved high consistency but faced challenges with entity recognition, especially in Hindi.

4.4 Models Misapply Gender in Multi-Entity Sentences

Gender interference occurs when a model incorrectly alters words associated with all entities' genders instead of only the gendered terms in sentences with multiple human entities. To measure correct transformation of gendered terms, we use gender accuracy, which counts only the changes to the intended gendered words. To further penalize any modifications of non-gendered words, we introduce Gendered IoU (GIoU), which is a stricter metric that penalizes models for making unintended edits. This is illustrated by an example result of LLama family of models on the multiple entities cases in our Hindi dataset (Figure 8) of appendix. Thus, a large difference between gender accuracy and GIoU indicates that models often transform non-gendered terms and suffer from gender interference and limited instruction following capability for this task.

4.5 French: Complex Morphology Amplifies Bias and Challenges Pronoun Agreement

French's larger dataset and complex morphology diluted performance, amplifying training imbalances, a trend evident in the GIoU scores presented in Figure 5 of appendix. Larger models exhibited masculine bias, while smaller models struggled significantly. Possessive pronoun agreement (e.g., *son instructeur/son instructrice*), requiring possession-based gender disambiguation, posed challenges. Smaller models lacked the morphological understanding to handle this, whereas larger models performed more effectively, reflecting the impact of capacity on complex rule application.

4.6 Arabic: Lowest Scores with Persistent Masculine Bias in Plurals

Arabic's smaller, stricter dataset with intricate morphology yielded the lowest scores, as reflected in the GIoU scores in Figure 6 of appendix. Larger models mitigated complexity with balanced gender handling, while smaller models faltered, often showing masculine biases. Female plural agreement (e.g., *ka-mumaththilāt* for "actresses"), defaulting to masculine for female plural groups, highlighted inadequate training on gender-specific morphology, with most models over-applying masculine forms, even in all-female contexts.

4.7 Hindi: Feminine Skew and Entity Errors

Models achieved higher performance on the Hindi dataset of the MORPHOGEN benchmark, reflecting its simpler morphology with fewer gender nuances, as illustrated in the GIoU scores in Figure 7 of appendix. Larger models demonstrated superior performance with minimal gender disparity, while smaller models remained competitive, underscoring Hindi's accessibility. However, some models displayed a feminine bias in female-to-male conversions, and others showed weaker entity recognition due to erroneous gender modifications. Models in 8B–12B range exhibited stronger entity recognition abilities. Smaller models struggled on direct speech involving adjectives and occupations, and co-reference resolution (e.g., *śikṣak/śikṣikā* for "teacher") failing to resolve a speaker's gender, unlike larger models with robust co-reference handling.

5 Conclusions and Future Work

This paper introduced MORPHOGEN, a new multilingual benchmark for evaluating gender-aware morphological generation in LLMs, covering Hindi, French, and Arabic, three typologically diverse, gendered languages. MORPHOGEN focuses on a controlled first-person transformation task that isolates gender-sensitive morphological reasoning. We proposed novel evaluation metrics tailored to this setting and benchmarked 15 multilingual LLMs ranging from 2B to 70B parameters.

Our results show models often confuse gendered forms, especially with multiple entities, and exhibit biased masculine-to-feminine vs. feminine-to-masculine transformations, with some models showing strong directional bias. This highlights persistent limitations in LLMs' handling of gendered morphology. MORPHOGEN offers a foundation for studying morphological competence in multilingual models. Future work should expand it to include 2nd and 3rd person constructions, other gendered languages, and more complex discourse. Our work also enables developing gender-sensitive training and evaluating bias in generative tasks like translation, summarization, and dialogue.

References

[1] Arnav Goel, Medha Hira, Avinash Anand, Siddhesh Bangar, and Rajiv Ratn Shah. Advancements in scientific controllable text generation methods, 2023.

[2] Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. Multilingual large language model: A survey of resources, taxonomy and frontiers, 2024.

[3] Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. A survey on multilingual large language models: corpora, alignment, and bias. *Frontiers of Computer Science*, 19(11), April 2025.

[4] Kaiyu Huang, Fengran Mo, Xinyu Zhang, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, Jinan Xu, Jian-Yun Nie, and Yang Liu. A survey on large language models with multilingualism: Recent advances and new frontiers, 2025.

[5] Medha Hira, Arnav Goel, and Anubha Gupta. Crossvoice: Crosslingual prosody preserving cascade-s2st using transfer learning, 2024.

[6] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International conference on machine learning*, pages 4411–4421. PMLR, 2020.

[7] Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, et al. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *arXiv preprint arXiv:2412.03304*, 2024.

[8] Xu Huang, Wenhao Zhu, Hanxu Hu, Conghui He, Lei Li, Shujian Huang, and Fei Yuan. Benchmax: A comprehensive multilingual evaluation suite for large language models, 2025.

[9] Minghao Wu, Weixuan Wang, Sinuo Liu, Huifeng Yin, Xintong Wang, Yu Zhao, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. The bitter lesson learned from 2,000+ multilingual benchmarks, 2025.

[10] Andrea Piergentili, Beatrice Savoldi, Matteo Negri, and Luisa Bentivogli. Enhancing gender-inclusive machine translation with neomorphemes and large language models, 2024.

[11] Hila Gonen, Yova Kementchedjhieva, and Yoav Goldberg. How does grammatical gender affect noun representations in gender-marking languages?, 2019.

[12] Sunayana Sitaram, Adrian de Wynter, Isobel McCrum, Qilong Gu, and Si-Qing Chen. A multilingual, culture-first approach to addressing misgendering in llm applications. *arXiv preprint arXiv:2503.20302*, 2025.

[13] Matúš Pikuliak, Andrea Hrckova, Stefan Oresko, and Marián Šimko. Women are beautiful, men are leaders: Gender stereotypes in machine translation and language modeling, 2024.

[14] Guijin Son, Dongkeun Yoon, Juyoung Suk, Javier Aula-Blasco, Mano Aslan, Vu Trong Kim, Shayekh Bin Islam, Jaume Prats-Cristià, Lucía Tormo-Bañuelos, and Seungone Kim. Mm-eval: A multilingual meta-evaluation benchmark for llm-as-a-judge and reward models, 2025.

[15] Hengyu Luo, Zihao Li, Joseph Attieh, Sawal Devkota, Ona de Gibert, Shaoxiong Ji, Peiqin Lin, Bhavani Sai Praneeth Varma Mantina, Ananda Sreenidhi, Raúl Vázquez, Mengjie Wang, Samea Yusofi, and Jörg Tiedemann. Gloteval: A test suite for massively multilingual evaluation of large language models, 2025.

[16] Nishat Raihan, Antonios Anastasopoulos, and Marcos Zampieri. mhumaneval–a multilingual benchmark to evaluate large language models for code generation. *arXiv preprint arXiv:2410.15037*, 2024.

[17] Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia Antonino Di Gangi, Roldano Cattoni, and Marco Turchi. Gender in danger? evaluating speech translation technology on the must-she corpus, 2020.

[18] Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. Evaluating gender bias in machine translation. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy, July 2019. Association for Computational Linguistics.

[19] Rishav Hada, Safiya Husain, Varun Gumma, Harshita Diddee, Aditya Yadavalli, Agrima Seth, Nidhi Kulkarni, Ujwal Gadiraju, Aditya Vashistha, Vivek Seshadri, and Kalika Bali. Akal badi ya bias: An exploratory study of gender bias in hindi language technology, 2024.

[20] Haneh Rhel and Dmitri Roussinov. Large language models and arabic content: A review, 2025.

[21] Argentina Anna Rescigno, Eva Vanmassenhove, Johanna Monti, and Andy Way. A case study of natural gender phenomena in translation: A comparison of google translate, bing microsoft translator and deepl for english to italian, french and spanish. *CEUR Workshop Proceedings*, 2769:62–90, 2020. Publisher Copyright: Copyright © 2020 for this paper by its authors.; 7th Italian Conference on Computational Linguistics, CLiC-it 2020 ; Conference date: 01-03-2021 Through 03-03-2021.

[22] Viktor Mihaylov and Aleksandar Shtedritski. What an elegant bridge: Multilingual llms are biased similarly in different languages, 2024.

[23] Arnav Goel, Medha Hira, and Anubha Gupta. Exploring multilingual unseen speaker emotion recognition: Leveraging co-attention cues in multitask learning, 2024.

[24] Huihan Li, Arnav Goel, Keyu He, and Xiang Ren. Attributing culture-conditioned generations to pretraining corpora, 2025.

[25] Arnav Goel, Medha Hira, and Anubha Gupta. Multilingual prosody transfer: Comparing supervised transfer learning, 2024.

[26] Minwoo Lee, Hyukhun Koh, Minsung Kim, and Kyomin Jung. Fine-grained gender control in machine translation with large language models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5416–5430, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[27] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[28] Jay Gala, Pranjal A Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, et al. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*, 2023.

[29] NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe,

Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation, 2022.

[30] P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *ACL Anthology*, pages 79–86, 2024.

[31] Anna Currey, Maria Nadejde, Raghavendra Reddy Pappagari, Mia Mayer, Stanislas Lauly, Xing Niu, Benjamin Hsu, and Georgiana Dinu. MT-GenEval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4287–4299, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[32] Beatrice Savoldi, Eleonora Cupin, Manjinder Thind, Anne Lauscher, Andrea Piergentili, Matteo Negri, and Luisa Bentivogli. mgente: A multilingual resource for gender-neutral language and translation, 2025.

[33] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[34] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[35] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[36] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[37] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

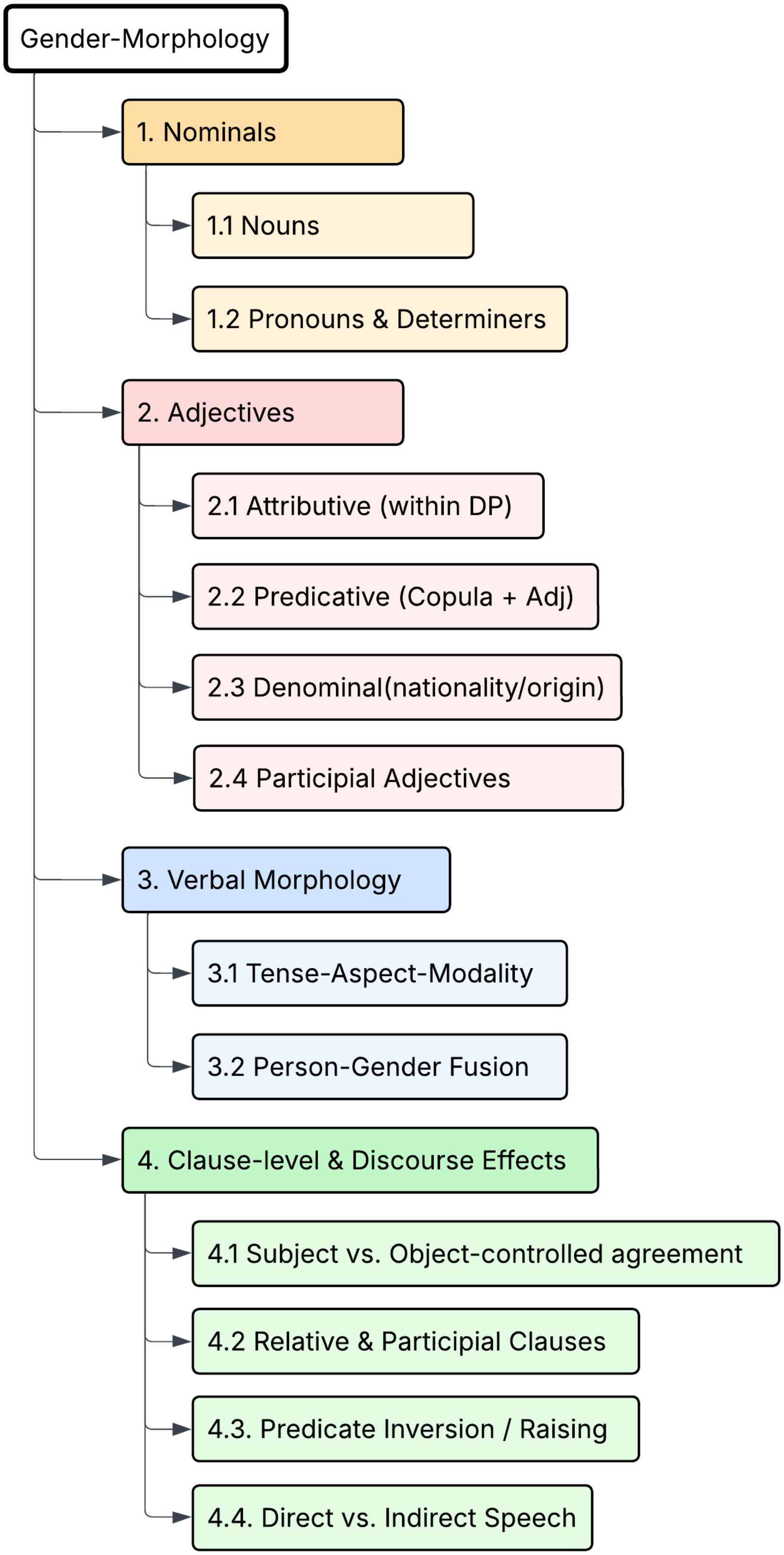
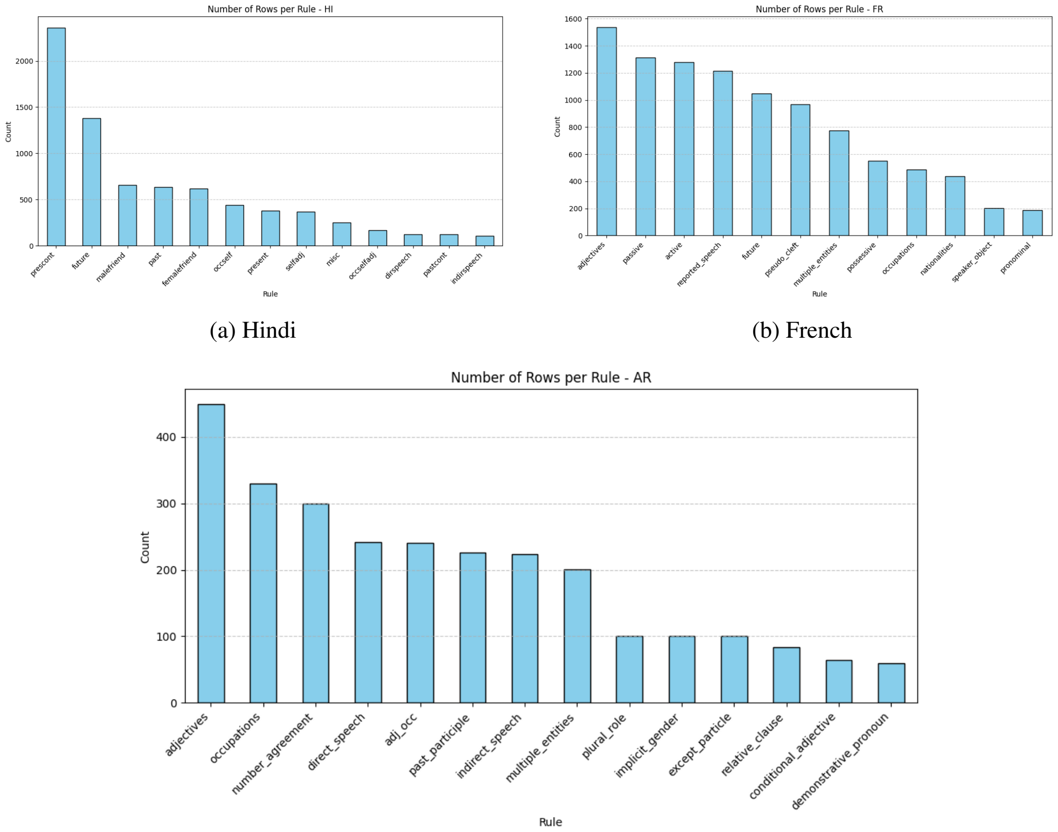

Gender-Morphology
1. Nominals
1.1 Nouns
1.2 Pronouns & Determiners
2. Adjectives
2.1 Attributive (within DP)
2.2 Predicative (Copula + Adj)
2.3 Denominal(nationality/origin)
2.4 Participial Adjectives
3. Verbal Morphology
3.1 Tense-Aspect-Modality
3.2 Person-Gender Fusion
4. Clause-level & Discourse Effects
4.1 Subject vs. Object-controlled agreement
4.2 Relative & Participial Clauses
4.3. Predicate Inversion / Raising
4.4. Direct vs. Indirect Speech

Figure 2: General morphological rules for grammatically gendered languages



(a) Hindi

(b) French

(c) Arabic

Figure 3: Distribution of Sentence Frequency Per Morphological Rule for Each Language

A Related Work

A.1 Existing Benchmarks on Multilingual LLMs

Recent advancements in multilingual LLM evaluation have produced several broad-coverage benchmarks. **XTREME** [6] emerged as a foundational multi-task benchmark spanning 40 languages and 9 tasks (e.g., NER, QA), though its focus on cross-lingual transfer left gaps in morphosyntactic evaluation. Subsequent works like **MM-Eval** [14] introduced meta-evaluation protocols for 18 languages, emphasizing multilingual consistency in LLM-as-judge scenarios, but remained task-agnostic to gender morphology. Resource-focused frameworks such as **GlotEval** [15] expanded coverage to hundreds of languages across seven NLP tasks, while **mHumanEval** [16] addressed code generation in 200+ languages via machine-translated prompts. Domain-specific efforts like **MuST-SHE** [17]. and **WinoMT** [18] pioneered gender-disambiguated MT datasets for Romance languages, though their narrow scope (1k examples per language) limited utility for LLM evaluation.

A.2 Evaluating Gendered Languages in Multilingual LLMs and NLP Systems

Grammatically gendered languages like French, Arabic, and Hindi pose unique evaluation challenges due to their morphological complexity. For Hindi, [19] revealed that LLMs struggle with gender-inflected verb conjugations and occupational noun morphology. Arabic evaluations [20] exposed performance gaps in dialectal gender agreement, while French analyses [21] demonstrated LLMs' tendency to default to masculine forms despite contextual cues.

Interestingly, a recent work [22] argues that none of the existing benchmarks systematically evaluate LLMs' application of gender morphology rules (e.g., adjective-noun concord) across diverse typologies, which is the gap addressed by our work. Other work looks at how culture and speech affects bias in LLM and NLP systems which further underscores the need to address this problem from multiple views and modalities [23, 24, 25].

B Dataset

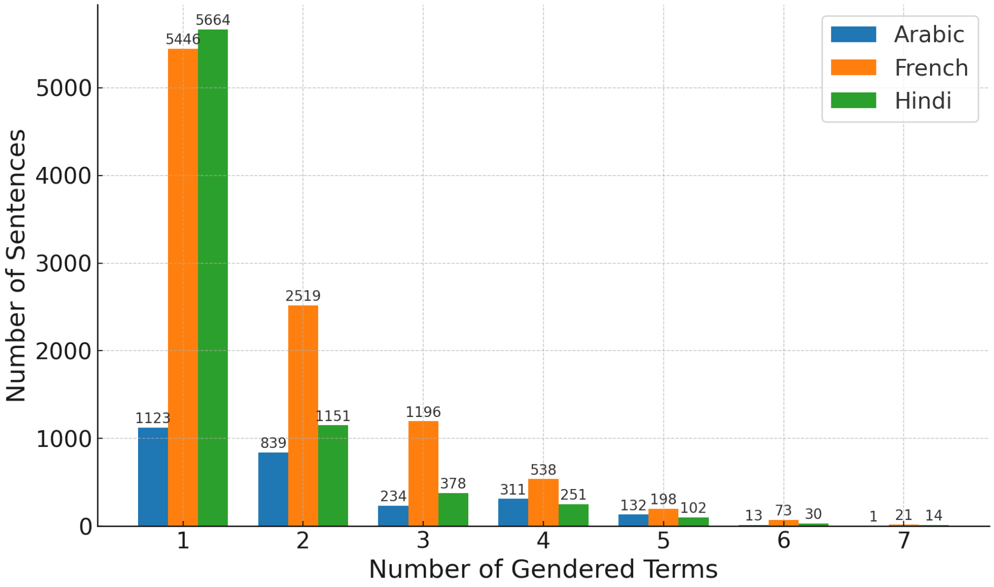


Figure 4: Gendered Terms Distribution in MORPHOGEN

B.1 Gender Morphology for Chosen Languages

MORPHOGEN comprises sentence pairs in three typologically diverse, grammatically gendered languages: **French**, **Arabic**, and **Hindi**. These were deliberately selected to capture a range of gender assignment strategies of semantic, morphological, and phonological nature offering a robust testbed for evaluating morphological generalization in multilingual LLMs. All three languages feature binary gender systems (masculine and feminine), but differ significantly in how gender is marked and propagated. This variation is depicted in Figure 1.

French combines semantic, morphological, and phonological cues. While suffixes like *-e* often indicate feminine gender, exceptions are common. Gender agreement is mandatory across determiners, adjectives, and verbs, but variability in marking makes it typologically distinct.

Arabic features a highly regular morphological system where gender is marked primarily via suffixation (e.g., *-a* for feminine). Agreement is strict and pervasive across verbs, adjectives, and pronouns, making it a consistent ground for evaluating morphological accuracy.

Hindi employs a natural gender system with partial morphological marking. Gender is semantically assigned, especially for animate nouns, and commonly marked via suffixes (e.g., *-ā* for masculine, *-ī* for feminine). Agreement extends to verbs, adjectives, and pronouns, but with moderate regularity due to exceptions.

Together, these languages exemplify distinct typological frameworks in gender morphology: French integrates phonological, morphological, and semantic gender assignment; Arabic employs regular morphological suffixation with strict agreement; and Hindi blends semantic natural gender with morphological suffixes.

B.2 Construction of Morphological Rules

To evaluate whether multilingual models can accurately perform gender transformation in first-person contexts, we construct a set of language-specific morphological rules grounded in linguistic theory, as

shown in Figure 3 of appendix. These rules are inspired by a general taxonomy of gender morphology across grammatically gendered languages (Figure 2) and are illustrated with concrete examples in Table 2 of Appendix. We present an overview for our motivation behind constructing these rules as below:
(1) Verbs and Tenses. Gender inflection on verbs depends on both tense and aspect, varying across languages. For instance, French present-tense verbs are gender-invariant, while past participles in compound tenses agree in gender with the subject. Our rules capture such tense-specific patterns.
(2) Adjectives and Role Nouns. Adjectives and identity-bearing nouns (e.g., occupations, nationalities) often mark speaker gender morphologically. We design transformation rules to reflect these regular and predictable gendered forms.
(3) Pronouns and Possessives. Gender marking in pronouns and possessives is language-dependent. Hindi marks the gender of the possessor, while French and Arabic express gender through grammatical agreement. Our rules reflect these alignment differences.
(4) Clause-Level Effects. Gender agreement may be influenced by sentence structure, especially in constructions involving passives or object-fronting. We include rules to account for such syntactic interactions that affect gender realization.
(5) Multiple Entities and Gender Interference. To evaluate a model's sensitivity to speaker identity, we introduce sentences with two human referents. Only the speaker's gender governs agreement, allowing us to test susceptibility to gender interference [26].

We provide detailed rules with examples for each language in the following tables in the Appendix: French (Tables 3, 4), Arabic (Table 5) and Hindi (Tables 6, 7).

B.3 Dataset Construction

We constructed the MORPHOGEN dataset capturing sentence-level gender transformations in French, Arabic, and Hindi through a structured pipeline grounded in linguistic principles. We began by identifying grammatical phenomena where a speaker's gender influences agreement or lexical choice, such as in tense and voice (e.g., active/passive), occupations and adjectives, pronouns and possessives, and multi-entity contexts prone to gender interference. For each case, we designed sentence templates (e.g., "I am a <occupation>") to ensure structural consistency and systematic coverage. Prompts specifying the rules, lexical arguments (e.g., occupation = doctor), and discourse contexts (e.g., politics, classroom, therapy) were used to generate English sentences via GPT-4o-mini [27]. These English sentences were translated into Hindi (using IndicTrans2 and GPT-4o-mini) [28, 27], Arabic (Grok-3)[4], and French (NLLB-200) [29], and then reviewed for accuracy by native speakers. Each sentence was manually corrected into both masculine and feminine forms by multiple annotators (aged 18–21 years), proficient in their respective languages. A total of 7 annotators participated. The resulting parallel gender-specific annotations form a high-quality gold-standard set for evaluating the model's sensitivity to morphosyntactic gender variation.

B.4 Comparison with Existing Datasets

Standard parallel corpora often default to masculine forms when gender is not explicitly marked. For instance, the EuroParl corpus includes speaker metadata but only 30% of its sentences are spoken by women, resulting in a male bias [30]. Such imbalance limits their suitability for evaluating gender accuracy. Specialized challenge sets exist but fall short for our speaker-gender restoration task:

(1) WinoMT targets occupational stereotypes across languages, including English–Hindi, but relies on rigid templates that models may overfit to [18]. **(2) MT-GenEval** improves diversity and realism for English–Hindi but lacks first-person sentences and speaker-gender labels [31]. **(3) MuST-SHE** offers speaker annotations and first-person content, but is not publicly available [17]. **(4) mGENTE** supports gender-neutral generation across languages [32], but lacks speaker-grounded, first-person constructions.

To our knowledge, no existing dataset:

1. Provides male and female translations for every sentence.
2. Aligns examples with grammatical triggers for gender inflection.

[4] https://x.ai/grok

3. Ensures balanced ground truth for both genders.
4. Covers the full spectrum of gender-marking phenomena.

Mining real transcripts is inefficient: most sentences are gender-neutral and few cover key structures. In contrast, prompting large language models under controlled templates enables efficient generation of diverse, balanced, and linguistically grounded examples across Hindi, Arabic, and French.

C Experimental Setup

C.1 Full Model Specifications

To effectively evaluate the performance of multilingual LLMs on MORPHOGEN, we conducted extensive benchmarking across 15 models spanning a diverse range of model families and parameter scales. The models evaluated include:

- **LLAMA**: LLAMA-3.1-8B, LLAMA-3.2-3B, LLAMA-3.3-70B [33]
- **Qwen**: Qwen3-4B, Qwen3-8B, Qwen3-14B, Qwen3-27B [34]
- **Gemma**: Gemma2-2B, Gemma2-9B, Gemma3-4B, Gemma3-12B, Gemma3-27B [35, 36]
- **Phi**: Phi4-14B [37]

Our goal was to cover a representative and practical spectrum of contemporary multilingual LLMs, ranging from lightweight models (e.g., 2B–4B parameters) suitable for deployment and industry use-cases, to high-capacity models (up to 70B parameters) that are expected to exhibit stronger multilingual generalization. These models were selected based on their widespread adoption, open-source availability, and explicit support for the three gendered languages under study.

C.2 Formal Metric Definitions

To evaluate model performance on the MORPHOGEN benchmark, we propose three complementary metrics that measure an LLM's ability to correctly perform gender-aware morphological transformations at different granularities. Note that for any sentence, we collect gendered terms by referring to its gender-counterfactual as presented in Section 2. The proposed metrics are defined as follows:

(1) Sentence-Level Gender Accuracy (SGA): This metric measures the proportion of correctly generated gendered terms in each sentence. For a given sentence, we compute the number of gendered words that were correctly modified (i.e., match the gold-standard target) and divide this by the total number of gendered terms in the reference sentence. The final score is the average of this ratio across all N sentences in the corpora:

$$\text{SGA} = \frac{1}{N} \sum_{i=1}^{N} \frac{|\text{Gendered}_i \cap \text{Mismatch}_i^c|}{|\text{Gendered}_i|}$$

As described in Section 2.2, the GENFORM task evaluates bidirectional gender transformation: masculine to feminine and vice versa. We report disaggregated results for each direction, denoted as SGA_M and SGA_F, corresponding to masculine-to-feminine and feminine-to-masculine conversions, respectively. Additionally, to evaluate any performance gaps between the masculine and feminine disaggregation, we report the gaps between the masculine and feminine scores $\triangle SGA$.

$$\triangle SGA = SGA_M - SGA_F$$

(2) Gender IoU Score (GIoU): Inspired by the Intersection-over-Union (IoU) metric commonly used in object detection, GIoU metric provides a stricter and more comprehensive measure of morphological transformation quality. It penalizes both over-generation (modifying non-gendered terms or incorrect gendered entities) and under-generation (failing to modify gendered terms). For each sentence, we computed the ratio between the no. of correctly transformed gendered terms to the union of gendered and mismatched terms. The final score is the mean of sentence-level IOU values:

$$\text{GIoU} = \frac{1}{N} \sum_{i=1}^{N} \frac{|\text{Gendered}_i \cap \text{Mismatch}_i^c|}{|\text{Gendered}_i \cup \text{Mismatch}_i|}$$

This metric captures both precision and recall and is especially useful in sentences with multiple entities or partial gender relevance, where models may hallucinate or overlook certain terms. Again, we report disaggregated results for each direction i.e., $GIoU_M$ and $GIoU_F$, corresponding to masculine-to-feminine and feminine-to-masculine conversions, respectively.

(3) Corpus-Level Gender Accuracy (CGA) This is a corpus-level aggregation of gender correctness. Instead of averaging per-sentence ratios, we computed the ratio of no. of correctly generated gendered terms across the entire test set to the total no. of reference gendered terms in the corpus. This provides a holistic measure of overall transformation quality at an *n*-gram level:

$$\text{CGA} = \frac{\sum_{i=1}^{N} |\text{Gendered}_i \cap \text{Mismatch}_i^c|}{\sum_{i=1}^{N} |\text{Gendered}_i|}$$

D Gender-Morphology

Different languages express grammatical gender through distinct morphological patterns. An overview of these patterns is shown in Figure 2, with illustrative examples in Table 2. These patterns motivate our focus on three gendered languages: French, Arabic, and Hindi.

For each of these languages, we provide example snippets along with the corresponding morphological rules in Tables 3, 4, 5, 6, and 7.

E Model Hyperparameters and Compute Used

For all models evaluated on the MORPHOGEN benchmark, we used a standardized inference configuration to ensure consistency across generations. The input prompt was constructed using the model-specific chat template, and all models were queried in a zero-shot setting without any few-shot examples.

Generation Parameters. We used the following generation hyperparameters for all models, unless otherwise noted:

- **Sampling Strategy:** Deterministic (no sampling)
- **do_sample:** `False`
- **Max New Tokens:** 256
- **Temperature:** 0.1 (low temperature for controlled and accurate generations)
- **Top-p:** 0.95
- **Top-k:** Not used (default)
- **Num Return Sequences:** 1
- **Batch Size for Inference:** 1 (due to varied token limits across models)

All generations were performed with:

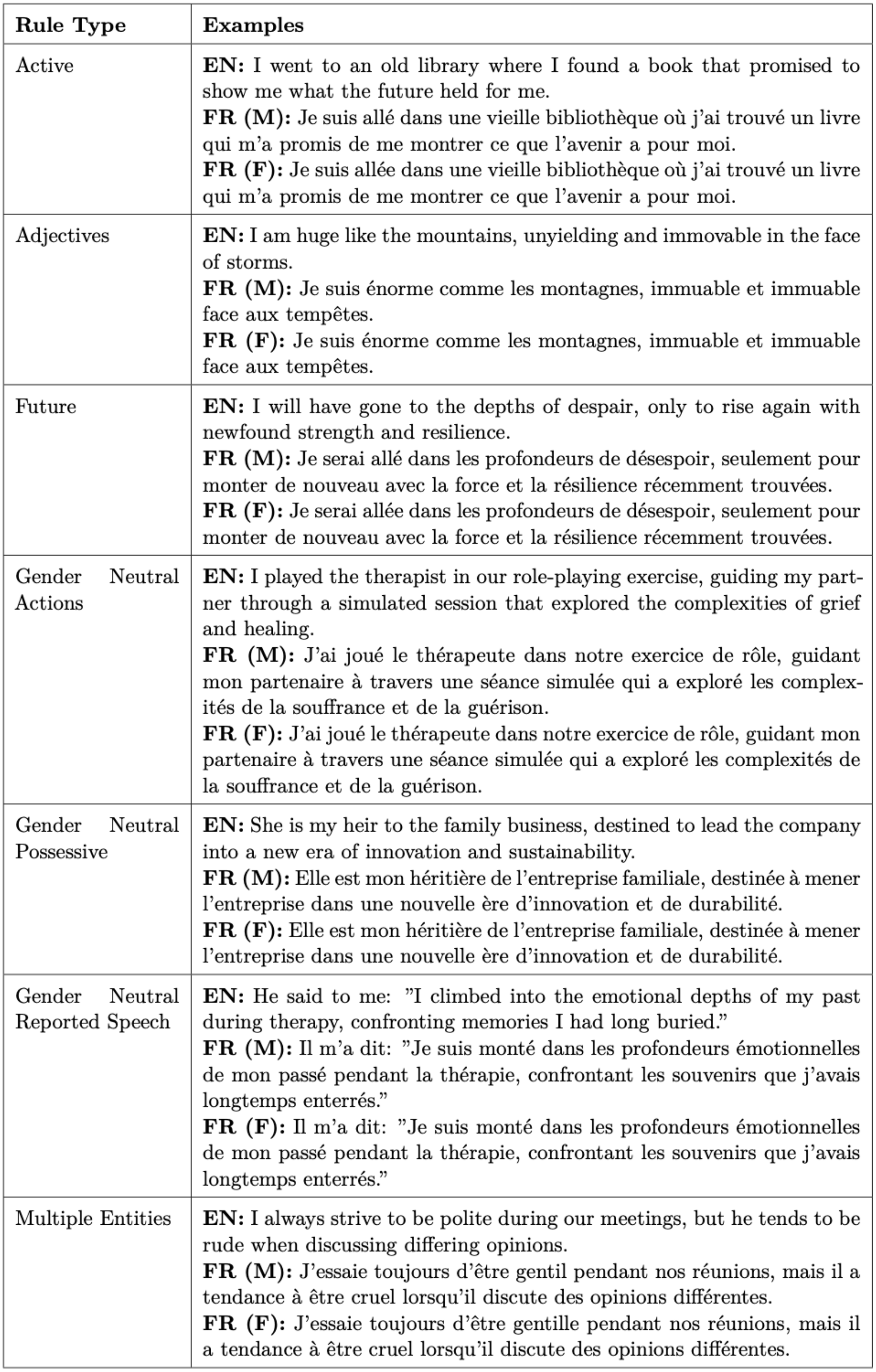
- **eos_token_id:** Set to the tokenizer's EOS token
- **pad_token_id:** Set to the tokenizer's PAD token if defined, else fallback to EOS

Compute Infrastructure. All experiments were run on an NVIDIA DGX A100 server equipped with 8 NVIDIA A100 GPUs, each with 40GB VRAM. While most models were executed using a single A100 GPU, larger models (e.g., mixture-of-experts or 65B+ parameter class) were distributed across multiple GPUs as needed via tensor or model parallelism.

This setup ensured sufficient compute headroom for large-scale inference and supported parallelized benchmarking across multiple languages and prompts.

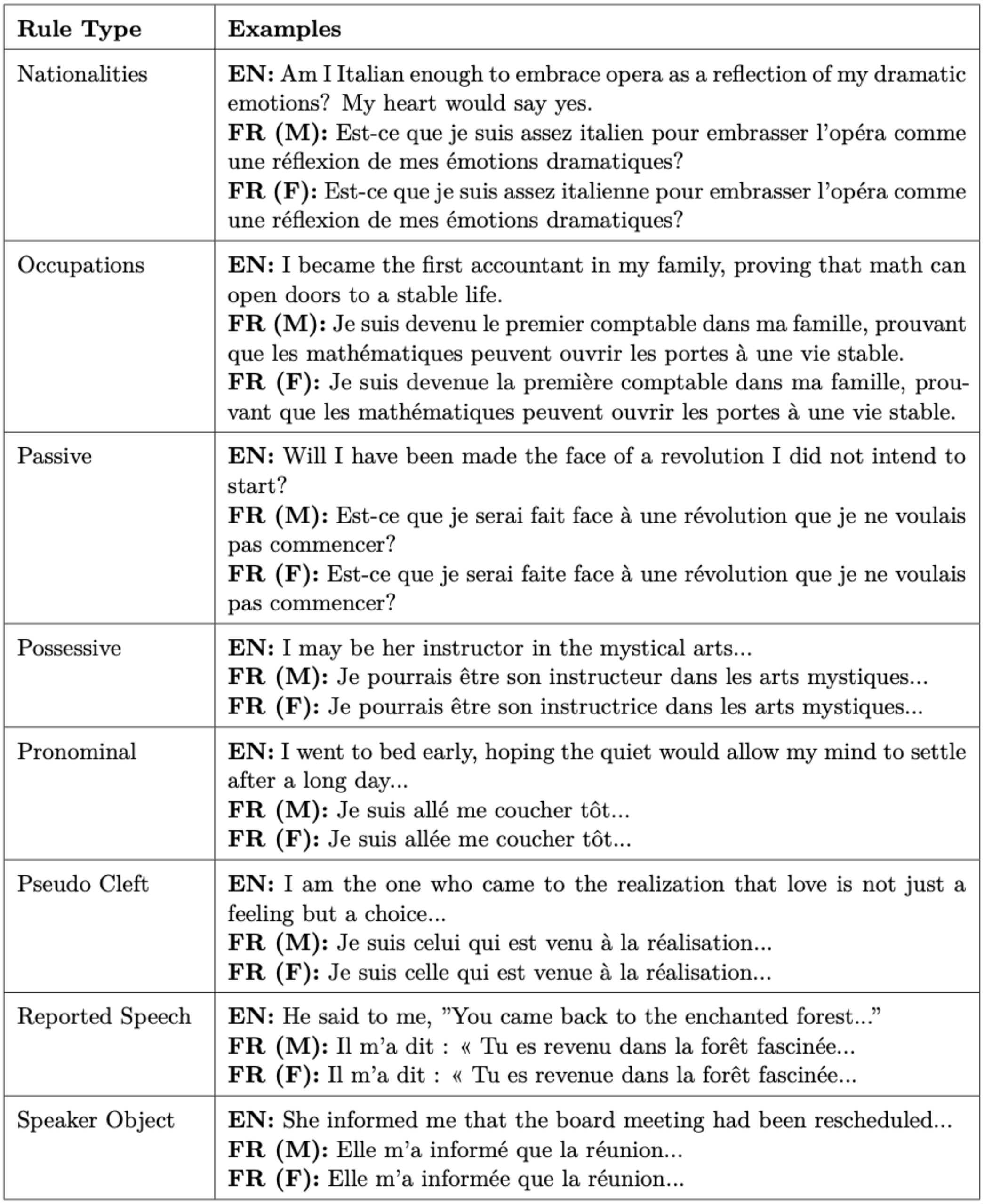
Category	**Subcategory**	**Details / Examples**
1. Nominals	1.1 Nouns	
	Epicene vs. Natural Gender	• Epicene: *la víctima* (Sp.) — always feminine • Natural: *el juez / la jueza* (Sp.)
	Role Nouns with Overt Gender	• *acteur / actrice* (Fr.) • *profesor / profesora* (Sp.)
	Appositive “Role As”	• Invariable: *en tant que médecin* (Fr.) • Relative: *el que fue médico / la que fue médica* (Sp.)
	1.2 Pronouns & Determiners	
	Personal Pronouns	• *he / she* (Eng.), *il / elle* (Fr.), *o* (Tur.) • Case-based gender (Slavic obliques)
	Possessives	• *mon / ma / mes* (Fr.) — agree with noun, not speaker • *мой / моя / моё* (Rus.) — speaker agreement in some contexts
	Demonstratives & Quantifiers	• *ten / ta / to* (Pol.), *každý / každá* (Czech)
2. Adjectives	2.1 Attributive (within DP)	• *une actrice italienne / un acteur italien* (Fr.)
	2.2 Predicative (Copula + Adj)	• *Él es mexicano / Ella es mexicana* (Sp.)
	2.3 Denominal (Nationality/Origin)	• *Je suis français / Je suis française* (Fr.) • *Sono inglese* (It.) — invariant
	2.4 Participial Adjectives	• *cansado / cansada* (Sp.), *allé / allée* (Fr.)
3. Verbal Morphology	3.1 Tense-Aspect-Modality	
	Past / Perfective	• *пришёл / пришла* (Rus.) • *khāyā / khāyī* (Hi.)
	Progressive / Continuous	• *je suis en train d’écrire* (Fr.) — no gender on gerund
	Future	• *sa-yaktubu* (Ar.) — prefix only
	Mood & Voice	• Active vs. Passive: – *Elle est aimée / Il est aimé* (Fr.) – *khāyā gayā / khāyī gayī* (Hi.) • Causative & Reflexive: generally mirror active agreement
	3.2 Person-Gender Fusion	• *katabtu / katabti* “I wrote” (Ar.) • *пришёл / пришла* for “I” (Slavic past tense)
4. Clause-Level & Discourse	4.1 Subject vs. Object Agreement	• *Je suis allé(e)* — PP agrees with subject • *Il m’a tué(e)* — PP agrees with object (Fr.)
	4.2 Relative & Participial Clauses	• *Soy el que fue médico / la que fue médica* (Sp.)
	4.3 Predicate Inversion / Raising	• *En tant que ingénieur(e)*
	4.4 Direct vs. Indirect Speech	• Direct: *Il m’a dit : « Je suis grand(e) »* — third person’s gender • Indirect: *Il m’a dit que j’étais grand(e)* — first person’s gender

Table 2: Gender Morphology Overview



Rule Type	Examples
Active	**EN:** I went to an old library where I found a book that promised to show me what the future held for me. **FR (M):** Je suis allé dans une vieille bibliothèque où j'ai trouvé un livre qui m'a promis de me montrer ce que l'avenir a pour moi. **FR (F):** Je suis allée dans une vieille bibliothèque où j'ai trouvé un livre qui m'a promis de me montrer ce que l'avenir a pour moi.
Adjectives	**EN:** I am huge like the mountains, unyielding and immovable in the face of storms. **FR (M):** Je suis énorme comme les montagnes, immuable et immuable face aux tempêtes. **FR (F):** Je suis énorme comme les montagnes, immuable et immuable face aux tempêtes.
Future	**EN:** I will have gone to the depths of despair, only to rise again with newfound strength and resilience. **FR (M):** Je serai allé dans les profondeurs de désespoir, seulement pour monter de nouveau avec la force et la résilience récemment trouvées. **FR (F):** Je serai allée dans les profondeurs de désespoir, seulement pour monter de nouveau avec la force et la résilience récemment trouvées.
Gender Neutral Actions	**EN:** I played the therapist in our role-playing exercise, guiding my partner through a simulated session that explored the complexities of grief and healing. **FR (M):** J'ai joué le thérapeute dans notre exercice de rôle, guidant mon partenaire à travers une séance simulée qui a exploré les complexités de la souffrance et de la guérison. **FR (F):** J'ai joué le thérapeute dans notre exercice de rôle, guidant mon partenaire à travers une séance simulée qui a exploré les complexités de la souffrance et de la guérison.
Gender Neutral Possessive	**EN:** She is my heir to the family business, destined to lead the company into a new era of innovation and sustainability. **FR (M):** Elle est mon héritière de l'entreprise familiale, destinée à mener l'entreprise dans une nouvelle ère d'innovation et de durabilité. **FR (F):** Elle est mon héritière de l'entreprise familiale, destinée à mener l'entreprise dans une nouvelle ère d'innovation et de durabilité.
Gender Neutral Reported Speech	**EN:** He said to me: ”I climbed into the emotional depths of my past during therapy, confronting memories I had long buried.” **FR (M):** Il m'a dit: ”Je suis monté dans les profondeurs émotionnelles de mon passé pendant la thérapie, confrontant les souvenirs que j'avais longtemps enterrés.” **FR (F):** Il m'a dit: ”Je suis monté dans les profondeurs émotionnelles de mon passé pendant la thérapie, confrontant les souvenirs que j'avais longtemps enterrés.”
Multiple Entities	**EN:** I always strive to be polite during our meetings, but he tends to be rude when discussing differing opinions. **FR (M):** J'essaie toujours d'être gentil pendant nos réunions, mais il a tendance à être cruel lorsqu'il discute des opinions différentes. **FR (F):** J'essaie toujours d'être gentille pendant nos réunions, mais il a tendance à être cruel lorsqu'il discute des opinions différentes.

Table 3: French Gendered Grammar Examples Across Rule Types [1]



Rule Type	**Examples**
Nationalities	**EN:** Am I Italian enough to embrace opera as a reflection of my dramatic emotions? My heart would say yes. **FR (M):** Est-ce que je suis assez italien pour embrasser l'opéra comme une réflexion de mes émotions dramatiques? **FR (F):** Est-ce que je suis assez italienne pour embrasser l'opéra comme une réflexion de mes émotions dramatiques?
Occupations	**EN:** I became the first accountant in my family, proving that math can open doors to a stable life. **FR (M):** Je suis devenu le premier comptable dans ma famille, prouvant que les mathématiques peuvent ouvrir les portes à une vie stable. **FR (F):** Je suis devenue la première comptable dans ma famille, prouvant que les mathématiques peuvent ouvrir les portes à une vie stable.
Passive	**EN:** Will I have been made the face of a revolution I did not intend to start? **FR (M):** Est-ce que je serai fait face à une révolution que je ne voulais pas commencer? **FR (F):** Est-ce que je serai faite face à une révolution que je ne voulais pas commencer?
Possessive	**EN:** I may be her instructor in the mystical arts... **FR (M):** Je pourrais être son instructeur dans les arts mystiques... **FR (F):** Je pourrais être son instructrice dans les arts mystiques...
Pronominal	**EN:** I went to bed early, hoping the quiet would allow my mind to settle after a long day... **FR (M):** Je suis allé me coucher tôt... **FR (F):** Je suis allée me coucher tôt...
Pseudo Cleft	**EN:** I am the one who came to the realization that love is not just a feeling but a choice... **FR (M):** Je suis celui qui est venu à la réalisation... **FR (F):** Je suis celle qui est venue à la réalisation...
Reported Speech	**EN:** He said to me, ”You came back to the enchanted forest...” **FR (M):** Il m'a dit : « Tu es revenu dans la forêt fascinée... **FR (F):** Il m'a dit : « Tu es revenue dans la forêt fascinée...
Speaker Object	**EN:** She informed me that the board meeting had been rescheduled... **FR (M):** Elle m'a informé que la réunion... **FR (F):** Elle m'a informée que la réunion...

Table 4: French Gendered Grammar Examples Across Rule Types [2]

Rule Type	**Examples**
Adjective	I am a responsible person who organizes community events. أنا شخص مسؤول أنظم فعاليات المجتمع. أنا شخصية مسؤولة أنظم فعاليات المجتمع.
Occupation	I am a doctor who treats patients in the clinic. أنا طبيب أعالج المرضى في العيادة. أنا طبيبة أعالج المرضى في العيادة.
Adjective + Occupation	I am a dedicated teacher who inspires students daily. أنا معلم مخلص ألهم الطلاب يومياً. أنا معلمة مخلصة ألهم الطلاب يومياً.
Plural Form	My sister and I, as actresses, performed in the play. أختي وأنا، كممثلين، قدمنا في المسرحية. أختي وأنا، كممثلات، قدمنا في المسرحية.
Relative Clause	I am the one who, as a poet, wrote the award-winning verse. أنا الذي، كشاعر، كتب الأبيات الفائزة بالجائزة. أنا التي، كشاعرة، كتبت الأبيات الفائزة بالجائزة.
Number Agreement	My colleague and I, as researchers, conducted the experiment. زميلي وأنا، كباحثين، أجرينا التجربة. زميلتي وأنا، كباحثات، أجرينا التجربة.
Demonstrative Pronoun	This is me, a teacher, guiding my students to become the leaders of tomorrow. هذا أنا، معلم، أهدي طلابي ليصبحوا قادة الغد. هذه أنا، معلمة، أهدي طالباتي ليصبحن قائدات الغد.
Conditional Adjective	If I were confident, I would lead the project as a leader. لو كنت واثقًا، لقدت المشروع كقائد. لو كنت واثقة، لقدت المشروع كقائدة.
Multiple Entity	Among the team, I stood out as the only engineer who solved the problem. بين الفريق، برزت كالمهندس الوحيد الذي حل المشكلة. بين الفريق، برزت كالمهندسة الوحيدة التي حلت المشكلة.
Implicit Gender Cue	I stood alone in the meeting, presenting my ideas confidently. وقفت وحدي في الاجتماع، أقدم أفكاري بثقة. وقفت وحدي في الاجتماع، أقدم أفكاري بثقة.
Direct Speech	I said, "I am a poet who crafts verses of hope." قلت: "أنا شاعر أصوغ أبيات الأمل." قلت: "أنا شاعرة أصوغ أبيات الأمل."
Indirect Speech	I told them that I was a consultant who advised the company. أخبرتهم أنني مستشار نصحت الشركة. أخبرتهم أنني مستشارة نصحت الشركة.
Past Participle	I have been a teacher who designed engaging history lessons. كنت معلمًا صممت دروس تاريخ جذابة. كنت معلمة صممت دروس تاريخ جذابة.
Exception Particle	No one attended the lecture except me, the student. ما حضر أحد المحاضرة إلا أنا الطالب. ما حضرت أحد المحاضرة إلا أنا الطالبة.

Table 5: Arabic Gendered Grammar Examples Across Rule Types

Rule Type	**Examples**
Adjectives	I was so tall that I had to duck under the doorframe every time I entered the classroom. **(M):** मैं इतना लंबा था कि मुझे हर बार कक्षा में प्रवेश करते समय दरवाजे के फ्रेम के नीचे झुकना पड़ता था। **(F):** मैं इतनी लंबी थी कि मुझे हर बार कक्षा में प्रवेश करते समय दरवाजे के फ्रेम के नीचे झुकना पड़ता था।
Direct Speech (Adj + Occupation)	”My teacher told me, ””You are a brilliant student.””” **(M):** मेरे शिक्षक ने मुझसे कहा, ""तुम एक शानदार छात्र हो।"" **(F):** मेरे शिक्षक ने मुझसे कहा, ""तुम एक शानदार छात्रा हो।""
Direct Speech (Complex)	”She said to me, ””You are a lovely person who brings joy to those around you...””” **(M):** उसने मुझसे कहा, ""तुम एक प्यारा व्यक्ति हो जो अपने आसपास के लोगों को खुशी देता है..."" **(F):** उसने मुझसे कहा, ""तुम एक प्यारी व्यक्ति हो जो अपने आसपास के लोगों को खुशी देती है...""
Direct Speech (Conditional)	”He said to me, ””You had already left the party when I arrived.””” **(M):** उसने मुझसे कहा, ""तू पार्टी से पहले ही चला गया था, जब मैं पहुँचा।"" **(F):** उसने मुझसे कहा, ""तू पार्टी से पहले ही चली गई थी, जब मैं पहुँचा।""
Female Friend	I know her from time to time, my friend works as a financial manager. **(M):** मैं उसे काफ़ी समय से जानता हूँ, मेरी दोस्त वित्तीय प्रबंधक का काम करती है। **(F):** मैं उसे काफ़ी समय से जानती हूँ, मेरी दोस्त वित्तीय प्रबंधक का काम करती है।
Future	I will handle the issue personally. **(M):** मैं इस मुद्दे को व्यक्तिगत रूप से संभालूंगा। **(F):** मैं इस मुद्दे को व्यक्तिगत रूप से संभालूँगी।
Indirect Speech (Adj + Occupation)	My teacher told us that she was a passionate poet who found inspiration in nature. **(M/F):** मेरे शिक्षक ने हमें बताया कि वह एक उत्साही कवयित्री है जो प्रकृति से प्रेरणा पाती है।
Indirect Speech (Complex)	The doctor advised that I should be a good student who eats healthy food to stay fit. **(M):** डॉक्टर ने सलाह दी कि मुझे एक अच्छा छात्र होना चाहिए जो स्वस्थ भोजन खाता है ताकि फिट रहे। **(F):** डॉक्टर ने सलाह दी कि मुझे एक अच्छी छात्रा होनी चाहिए जो स्वस्थ भोजन खाती है ताकि फिट रहे।
Indirect Speech (Conditional)	My friend said to me that he had traveled to the mountains and enjoyed the fresh air. **(M/F):** मेरे दोस्त ने मुझसे कहा कि उसने पहाड़ों की यात्रा करके ताज़ी हवा का आनंद लिया था।

Table 6: Hindi Gendered Grammar Examples Across Rule Types [1]

Rule Type	**Examples**
Male Friend	I know him from time to time, my friend works as a marketing specialist. **(M):** मैं उसे काफ़ी समय से जानता हूँ, मेरा दोस्त विपणन विशेषज्ञ का काम करता है। **(F):** मैं उसे काफ़ी समय से जानती हूँ, मेरा दोस्त विपणन विशेषज्ञ का काम करता है।
Neutral (मैंने/मुझे)	I need to review the effectiveness of the current emergency response plan. **(M/F):** मुझे वर्तमान आपातकालीन प्रतिक्रिया योजना की प्रभावशीलता की समीक्षा करनी है।
Neutral Occupation	As a nurse, I managed the care of wounded soldiers in the field hospital. **(M):** एक नर्स के रूप में मैंने युद्ध क्षेत्र के अस्पताल में घायल सैनिकों की देखभाल करता था। **(F):** एक नर्स के रूप में मैंने युद्ध क्षेत्र के अस्पताल में घायल सैनिकों की देखभाल करती थी।
Neutral Present-Past	I am committed to ensuring that every soldier is respected. **(M/F):** मैं यह सुनिश्चित करने के लिए प्रतिबद्ध हूं कि प्रत्येक सैनिक का सम्मान किया जाए।
Occupations	I was a passionate writer who loved crafting stories for my students. **(M):** मैं एक उत्साही लेखक था जो अपने छात्रों के लिए कहानियाँ लिखना पसंद करता था। **(F):** मैं एक उत्साही लेखिका थी जो अपने छात्रों के लिए कहानियाँ लिखना पसंद करती थी।
Occupations + Adjectives	People called me a curious good child... **(M):** लोगों ने मुझे एक जिज्ञासु अच्छे बच्चे के रूप में बुलाया... **(F):** लोगों ने मुझे एक जिज्ञासु अच्छी बच्ची के रूप में बुलाया...
Past	I was involved in the planning of the Salt Satyagraha. **(M):** मैं नमक सत्याग्रह की योजना बनाने में शामिल था। **(F):** मैं नमक सत्याग्रह की योजना बनाने में शामिल थी।
Past Continuous	I was preparing the legal arguments. **(M):** मैं कानूनी तर्क तैयार कर रहा था। **(F):** मैं कानूनी तर्क तैयार कर रही थी।
Present	I am scared of failing, so I avoid trying new things. **(M):** मैं असफल होने से डरता हूँ, इसलिए मैं नई चीज़ें आज़माने से बचता हूँ। **(F):** मैं असफल होने से डरती हूँ, इसलिए मैं नई चीज़ें आज़माने से बचती हूँ।
Present Continuous	I am preparing a study plan for you. **(M):** मैं आपके लिए एक अध्ययन योजना तैयार कर रहा हूँ। **(F):** मैं आपके लिए एक अध्ययन योजना तैयार कर रही हूँ।
TED	I apologize for any misunderstanding. **(M):** किसी भी गलतफहमी के लिए मैं माफी मांगता हूँ। **(F):** किसी भी गलतफहमी के लिए मैं माफी मांगती हूँ।
Third Person	Can you provide the exact time when the issue first occurred? **(M/F):** क्या आप सही समय बता सकते हैं कि समस्या पहली बार कब हुई थी?

Table 7: Hindi Gendered Grammar Examples Across Rule Types [2]

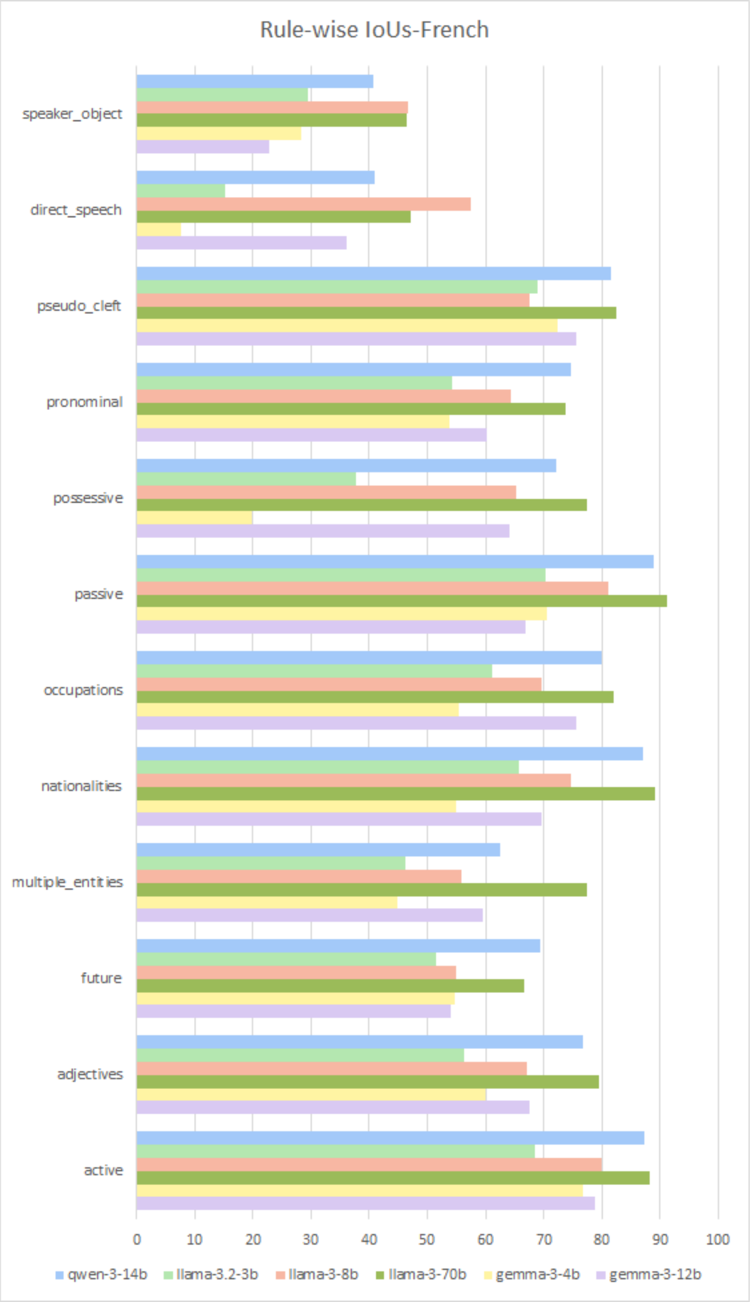


Figure 5: Rule based and model wise IoU metrics for French

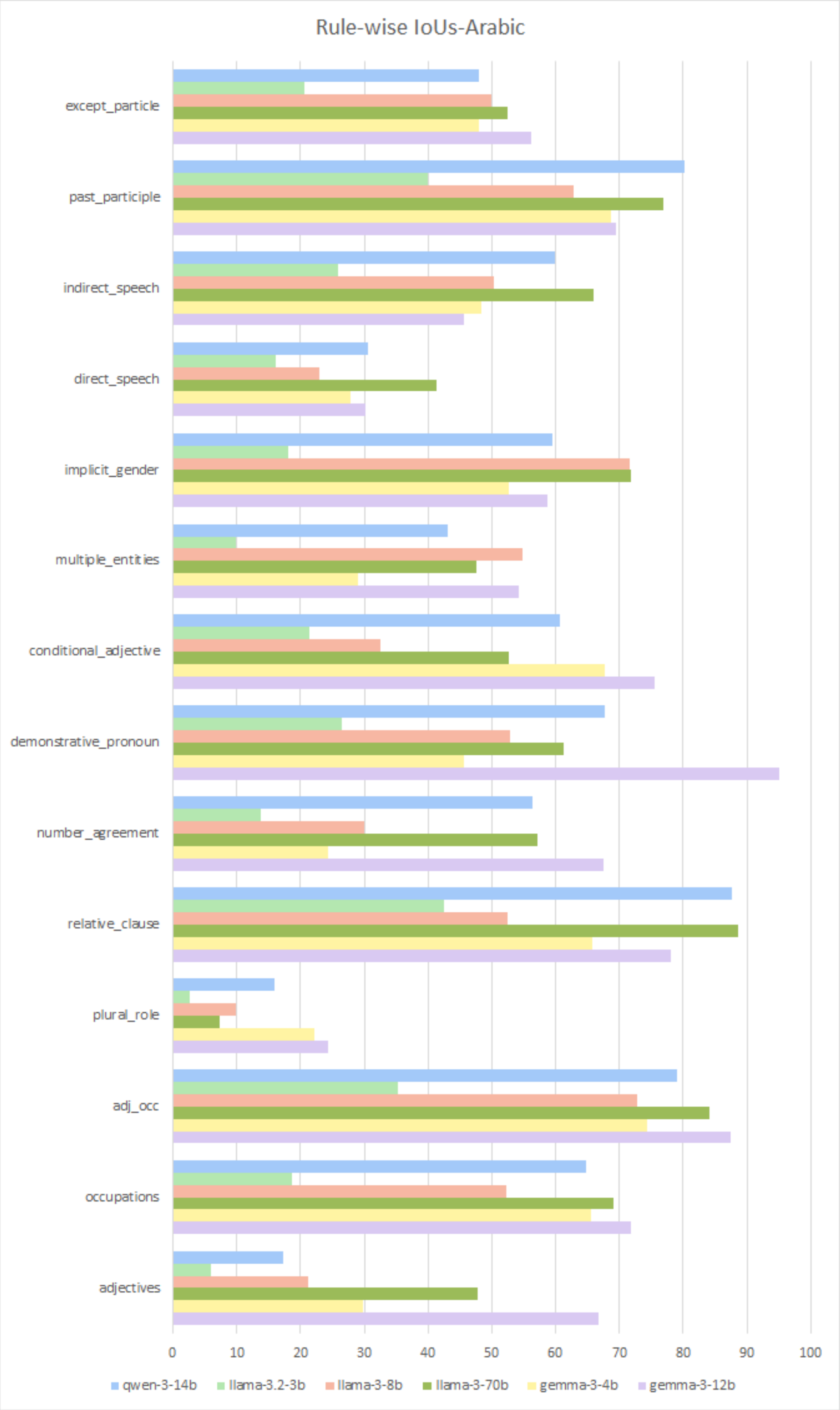


Figure 6: Rule based and model wise IoU metrics for Arabic

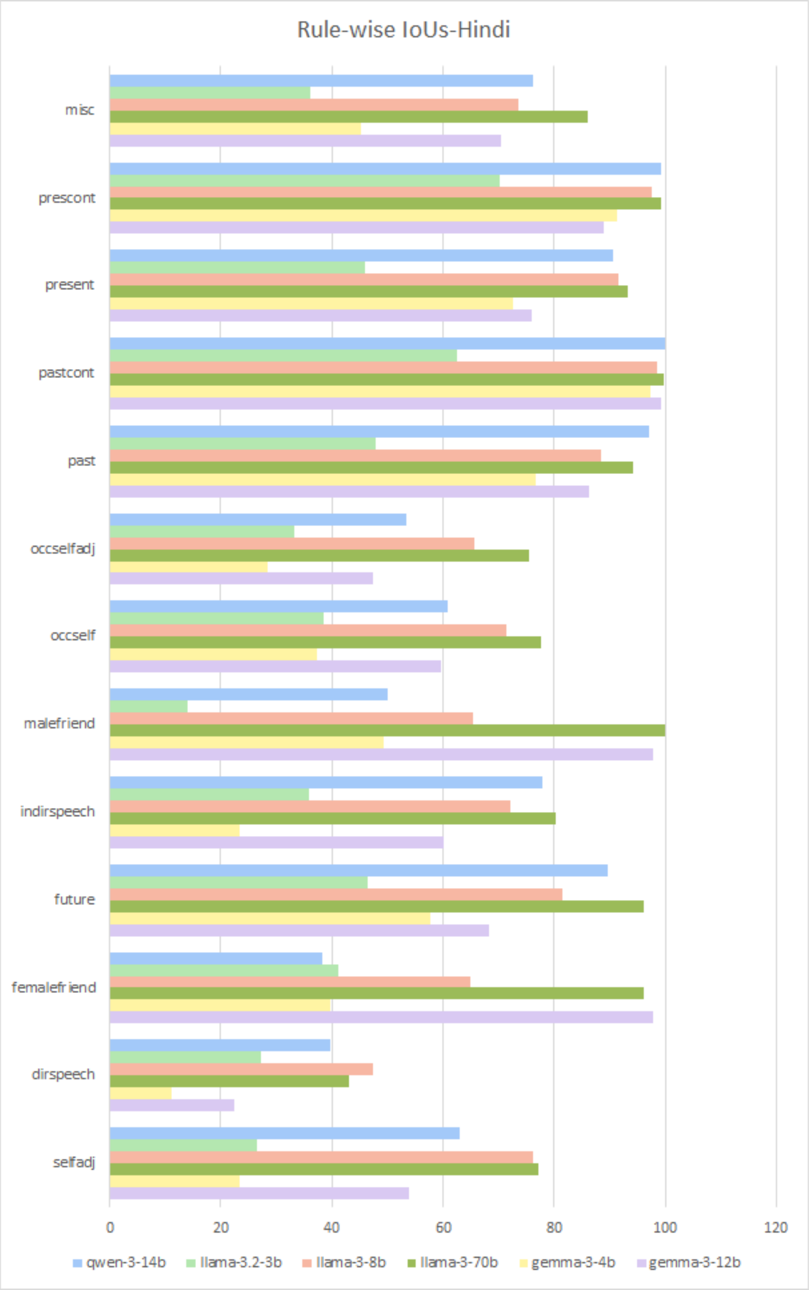
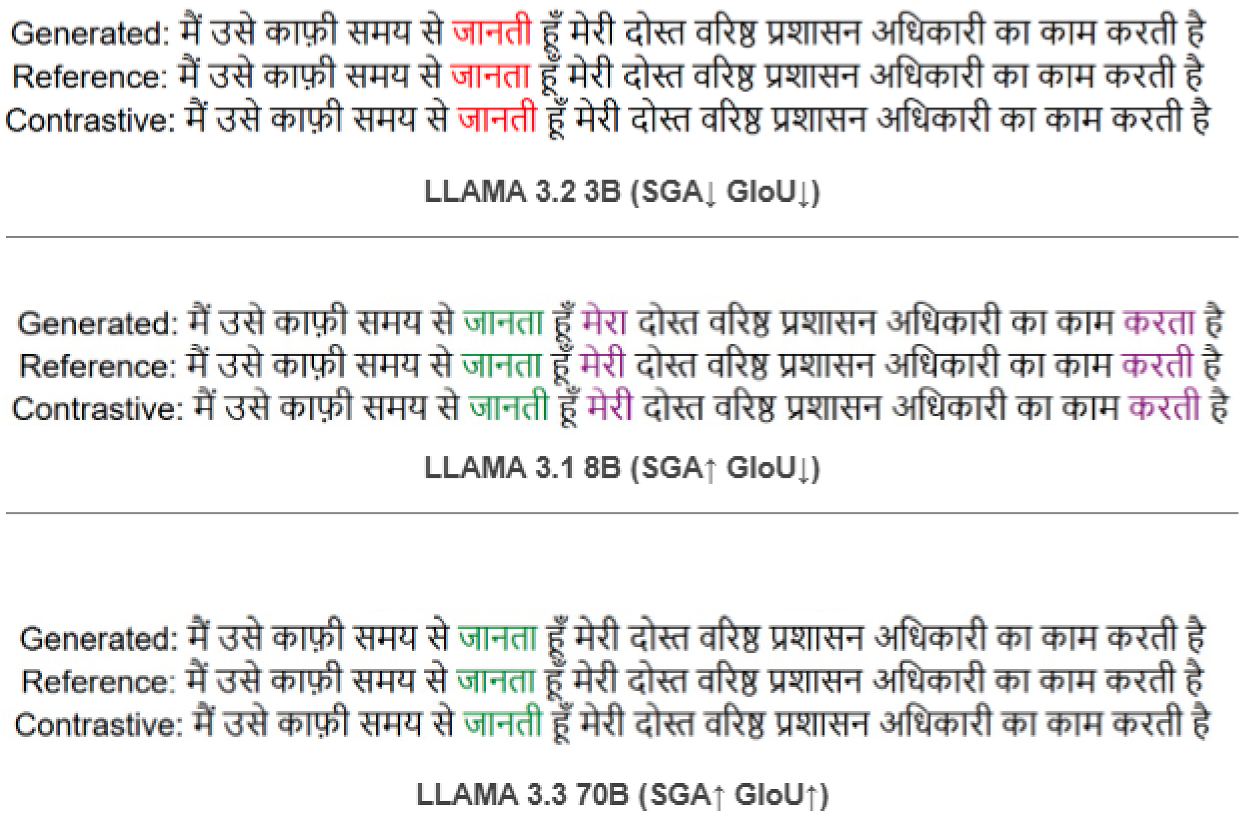


Figure 7: Rule based and model wise IoU metrics for Hindi



Generated: मैं उसे काफ़ी समय से जानती हूँ मेरी दोस्त वरिष्ठ प्रशासन अधिकारी का काम करती है
Reference: मैं उसे काफ़ी समय से जानता हूँ मेरी दोस्त वरिष्ठ प्रशासन अधिकारी का काम करती है
Contrastive: मैं उसे काफ़ी समय से जानती हूँ मेरी दोस्त वरिष्ठ प्रशासन अधिकारी का काम करती है

LLAMA 3.2 3B (SGA↓ GIoU↓)

Generated: मैं उसे काफ़ी समय से जानता हूँ मेरा दोस्त वरिष्ठ प्रशासन अधिकारी का काम करता है
Reference: मैं उसे काफ़ी समय से जानता हूँ मेरी दोस्त वरिष्ठ प्रशासन अधिकारी का काम करती है
Contrastive: मैं उसे काफ़ी समय से जानती हूँ मेरी दोस्त वरिष्ठ प्रशासन अधिकारी का काम करती है

LLAMA 3.1 8B (SGA↑ GIoU↓)

Generated: मैं उसे काफ़ी समय से जानता हूँ मेरी दोस्त वरिष्ठ प्रशासन अधिकारी का काम करती है
Reference: मैं उसे काफ़ी समय से जानता हूँ मेरी दोस्त वरिष्ठ प्रशासन अधिकारी का काम करती है
Contrastive: मैं उसे काफ़ी समय से जानती हूँ मेरी दोस्त वरिष्ठ प्रशासन अधिकारी का काम करती है

LLAMA 3.3 70B (SGA↑ GIoU↑)

Figure 8: Example of results of LLAMA family of models on multiple entities in Hindi Dataset

F Prompts

F.1 Sentence Generation

system_prompt = `"Suppose you are an Expert English Sentence Generating System."`

user_prompt =

```
Generate <Num_Sentences> English sentences.  Strictly adhere
to the format:  <Template>

Instructions:
1.  Only output the sentences--do not include any additional
text.
2.  Each sentence must be unique in its context and the nouns
used.
3.  Vary the sentence lengths and ensure they sound natural
and conversational.
4.  Use a variety of creative contexts, including but not
limited to [<Context_1>, <Context_2>, ..., <Context_n>].
```

The prompts are designed to guide a language model in generating diverse and natural-sounding English sentences. The system prompt establishes the model's role, while the user prompt provides clear, structured instructions to ensure variety, contextual relevance, and adherence to a specified format.

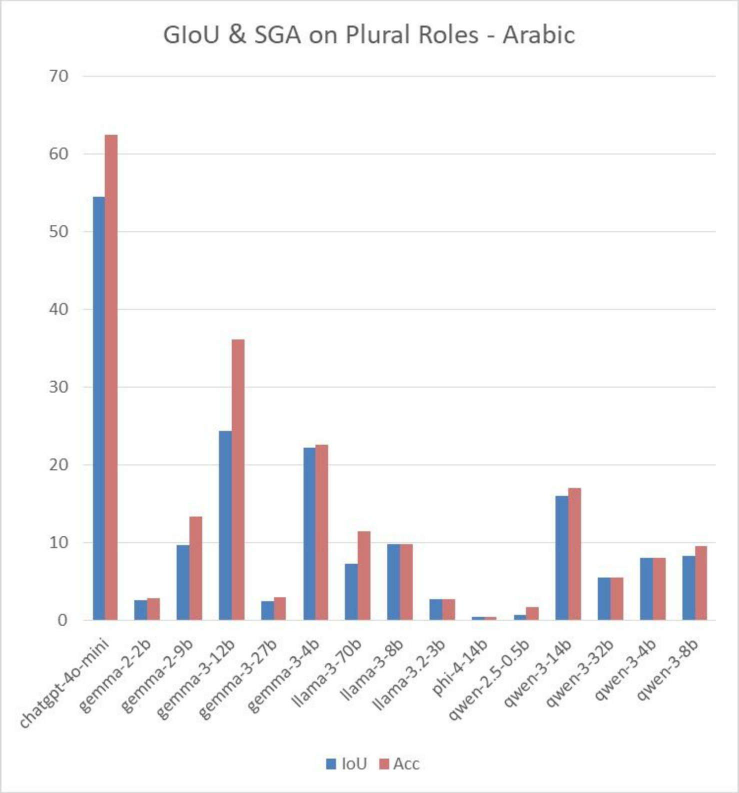


Figure 9: GIoU and SGA on Plural Roles in Arabic Dataset

F.2 Zero Shot Prompts

For zero-shot inference of the LLMs on the MORPHOGEN benchmark, we designed language-specific prompts to ensure precise gender-aware morphological transformations while preserving sentence structure. Although the prompts were provided to the models in the respective target languages (Hindi, French, or Arabic), the structure and content of the system and user prompts were consistent across languages.

The **system prompt** given to the model was:

> "You are a language assistant. Given a sentence in the target language and the gender of the speaker, adjust only the words that refer to the speaker to match the specified gender. Do not alter any other parts of the sentence. Return only the modified sentence with no explanations or extra words. If no change is required, return the sentence exactly as it is."

The **user prompt** provided the transformation instruction, depending on the speaker's gender:

- For male speakers: `Without changing the structure of the sentence, convert it as if it were spoken by a male speaker.`
- For female speakers: `Without changing the structure of the sentence, convert it as if it were spoken by a female speaker.`

This was followed by the sentence to be transformed: `Sentence to transform: [sentence].`

These prompts were designed to enforce minimal intervention, focusing solely on speaker-referring terms. This ensures the task evaluates the models' ability to perform gender-specific transformations without altering unrelated components of the sentence. The zero-shot setting tests the models' inherent linguistic knowledge, aligning with the benchmark's goal of assessing gender-aware morphological capabilities across diverse languages.

Model	$GIoU \uparrow$	$GIoU_M \uparrow$	$GIoU_F \uparrow$	$SGA \uparrow$	$SGA_M \uparrow$	$SGA_F \uparrow$	$\triangle SGA \downarrow$	$CGA \uparrow$
Qwen2.5-0.5B	0.35	0.69	0.05	0.35	0.69	0.05	0.63	0.21
Gemma2-2B	71.41	75.13	67.85	76.28	80.04	72.69	7.35	65.41
LLAMA-3.2-3B	48.54	19.85	76.90	53.08	20.57	85.22	**-64.65**	49.72
Gemma3-4B	67.50	60.29	73.21	71.75	63.76	78.08	-14.32	64.58
Qwen3-4B	62.84	61.96	63.70	73.74	75.41	72.08	3.33	68.51
LLAMA-3.1-8B	83.12	84.01	82.23	91.65	91.44	91.87	-0.43	89.21
Qwen3-8B	80.96	82.36	79.57	91.52	92.60	90.43	2.16	87.82
Gemma2-9B	85.47	82.78	88.20	87.42	83.78	91.12	-7.34	84.39
Gemma3-12B	79.91	75.69	84.16	84.93	80.48	89.41	-8.93	80.99
Phi4-14B	82.77	84.69	80.85	96.69	97.38	96.00	1.38	95.10
Qwen3-14B	80.68	83.62	77.74	90.22	94.12	86.30	7.81	85.80
Gemma3-27B	77.97	75.68	80.31	83.96	80.34	87.96	-7.61	82.56
Qwen3-32B	83.21	85.86	80.56	93.88	96.46	91.31	5.14	90.38
LLAMA-3.3-70B	93.33	95.04	91.62	94.06	95.89	92.22	3.67	91.40
GPT-4o-mini	88.81	90.08	87.54	95.73	96.04	95.42	0.62	93.36

Table 8: Performance metrics of different models on Hindi (% values; GIoU = Gender IoU, CGA = Corpus-Level Gender Accuracy, SGA = Sentence-Level Gender Accuracy, △SGA = Accuracy Gap, M = Male, F = Female)

Model	$GIoU \uparrow$	$GIoU_M \uparrow$	$GIoU_F \uparrow$	$SGA \uparrow$	$SGA_M \uparrow$	$SGA_F \uparrow$	$\triangle SGA \downarrow$	$CGA \uparrow$
Qwen2.5-0.5B	5.47	7.65	3.30	5.72	8.00	3.45	4.55	4.16
Gemma2-2B	39.73	37.29	42.16	40.90	38.32	43.47	-5.14	37.54
LLAMA-3.2-3B	54.49	60.19	48.85	59.20	64.94	53.52	11.42	53.48
Gemma3-4B	52.70	46.49	59.09	57.72	50.74	64.91	-14.16	51.60
Qwen3-4B	58.64	61.59	55.76	60.98	64.64	57.40	7.25	53.20
LLAMA-3.1-8B	67.89	70.67	64.62	82.75	84.44	80.75	3.69	81.76
Qwen3-8B	71.66	73.89	69.39	76.25	78.66	73.79	4.86	69.91
Gemma2-9B	60.52	62.02	59.02	65.48	66.11	64.84	1.26	55.56
Gemma3-12B	64.27	64.34	64.20	76.33	76.04	76.62	-0.58	74.26
Phi4-14B	79.84	81.46	78.22	89.68	90.26	89.09	1.17	87.70
Qwen3-14B	74.22	80.64	67.48	78.78	85.73	71.49	14.23	73.91
Gemma3-27B	71.89	75.47	68.11	83.11	86.78	79.25	7.53	79.63
Qwen3-32B	76.28	80.80	71.76	79.35	84.40	74.30	10.10	74.74
LLAMA-3.3-70B	76.68	83.53	69.81	80.76	88.33	73.17	15.15	76.08
GPT-4O-MINI	86.43	86.61	86.25	91.77	91.22	92.33	-1.11	90.27

Table 9: Performance metrics of different models on French (% values; GIoU = Gender IoU, CGA = Corpus-Level Gender Accuracy, SGA = Sentence-Level Gender Accuracy, △SGA = Accuracy Gap, M = Male, F = Female)

G Results

G.1 Examples of Outputs and Error Analysis

Figure 8 presents an error analysis of the LLaMA model family on a Hindi example involving gender and morphological agreement. Specifically, we compare the outputs of LLaMA 3.2 3B, LLaMA 3.1 8B, and LLaMA 3.3 70B.

- LLaMA 3.2 3B fails to produce the correct gendered forms, resulting in lower SGA and GIoU scores.

- LLaMA 3.1 8B correctly inflects for gender but incorrectly converts the first-person pronoun, which reduces its GIoU score.

- LLaMA 3.3 70B performs the transformation flawlessly, yielding high SGA and GIoU scores.

Model	$GIoU\uparrow$	$GIoU_M\uparrow$	$GIoU_F\uparrow$	$SGA\uparrow$	$SGA_M\uparrow$	$SGA_F\uparrow$	$\triangle SGA\downarrow$	$CGA\uparrow$
Qwen2.5-0.5B	4.14	7.31	0.72	6.28	10.37	1.88	8.49	4.59
Gemma2-2B	14.73	14.14	15.30	16.04	15.63	16.43	-0.81	14.10
LLAMA-3.2-3B	18.31	5.96	29.96	20.95	6.74	34.35	**-64.65**	17.75
Gemma3-4B	45.68	45.31	46.06	55.34	51.23	59.43	-8.20	48.93
Qwen3-4B	34.34	34.07	34.59	37.63	37.17	38.07	-0.90	35.97
LLAMA-3.1-8B	43.51	44.53	42.49	50.65	51.13	50.17	0.96	45.51
Qwen3-8B	45.89	47.93	43.89	51.44	53.99	48.96	5.03	51.01
Gemma2-9B	46.45	47.92	44.99	50.43	51.71	49.16	2.55	45.26
Gemma3-12B	62.76	64.69	60.82	69.37	70.62	68.12	2.50	65.52
Phi4-14B	57.08	62.24	52.20	66.51	69.89	63.31	6.58	66.15
Qwen3-14B	51.83	56.07	47.73	57.48	62.29	52.84	9.45	56.08
Gemma3-27B	70.33	71.47	69.19	77.12	76.70	77.53	-0.83	74.74
Qwen3-32B	50.69	57.32	44.10	56.57	62.56	50.62	11.94	53.00
LLAMA-3.3-70B	59.16	63.50	54.84	66.84	70.61	63.11	7.50	64.37
GPT-4o-mini	71.02	68.13	73.91	82.76	77.45	88.06	-10.61	80.27

Table 10: Performance metrics of different models on Arabic (% values; GIoU = Gender IoU, CGA = Corpus-Level Gender Accuracy, SGA = Sentence-Level Gender Accuracy, △SGA = Accuracy Gap, M = Male, F = Female)

H Limitations

This work presents MORPHOGEN, a large-scale, synthetic benchmark designed to evaluate multilingual language models on grammatical gender and morphological agreement across three typologically diverse and gendered languages: French, Arabic, and Hindi. While we believe MORPHOGEN represents an important step toward more inclusive and linguistically grounded evaluation of LLMs, several limitations remain.

First, the dataset currently covers only three languages, each represented in a standardized form without accounting for dialectal variation. French, Arabic, and Hindi each have dozens of dialects, many of which exhibit distinct grammatical and lexical gender patterns, which are not yet included in this release. **Second**, our Arabic dataset is smaller than the others, primarily due to limited availability of high-quality source data and fewer native Arabic-speaking annotators. **Third**, both Hindi and Arabic are predominantly binary-gendered languages; consequently, our current dataset focuses only on male and female speaker forms. We recognize this binary framing as a limitation and aim to extend the dataset to better represent gender as a spectrum in future work. Finally, while we also introduce multi-entity scenarios to evaluate gender interference, these are currently limited to two human referents per sentence. Expanding to more complex discourse scenarios with multiple gendered entities remains an important direction for future research.

Despite these limitations, MORPHOGEN provides a valuable and high-precision resource for advancing evaluation of how of LLMs across linguistically diverse settings.

I Ethical Considerations

Grammatical gender and morphological agreement are critical components of many NLP tasks, including machine translation, coreference resolution, and question answering. MORPHOGEN aims to address a key gap by providing a gender-focused evaluation benchmark for morphologically rich and typologically diverse languages such as French, Arabic, and Hindi. While the dataset is intended to advance fairness and inclusivity in multilingual language model development, we recognize several ethical considerations that arise from its creation and use.

First, the current task formulation is binary in nature, reflecting the masculine and feminine grammatical categories encoded in Hindi and Arabic. We acknowledge that this does not capture the full spectrum of gender identities and expressions. Future iterations of MORPHOGEN will seek to expand beyond binary gender, contingent on linguistic feasibility and community consultation. Additionally, French, Arabic, and Hindi encode gender distinctions that intersect with cultural, religious, and social norms—particularly through gendered references to animacy, occupations, or identities. To mitigate potential harm, we curated sentence prompts carefully to avoid reinforcing stereotypes and actively included constructions that challenge male-default biases (e.g., "doctor" or "leader" in feminine forms).

We also acknowledge the risk that gender-correct, grammatically coherent language could be misused to generate harmful content, including hate speech. While such misuse is outside the intended scope of this work, it remains a general concern in the release of any language resource. MORPHOGEN was designed with synthetic prompts and neutral scenarios to minimize these risks.

All annotations were conducted by undergraduate students aged 18–21. Annotators were compensated fairly and awarded certificates of recognition for their contributions. We ensured annotator well-being and did not include any sensitive, offensive, or personally identifiable content in the annotation tasks.

MORPHOGEN will be released under a Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0) license, which permits free use for research and non-commercial purposes. We strongly encourage the community to use this dataset to promote fairness and linguistic inclusivity in multilingual NLP systems.