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

Aditya Aggarwal[♡*] **Mehul Agarwal**[♡*] **Arnav Goel**[♡*]
Medha Hira[♡*] **Anubha Gupta**[♡†]

[♡]SBILab, Indraprastha Institute of Information Technology Delhi

anubha@iiitd.ac.in

Code Dataset

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 thus 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 NLP.

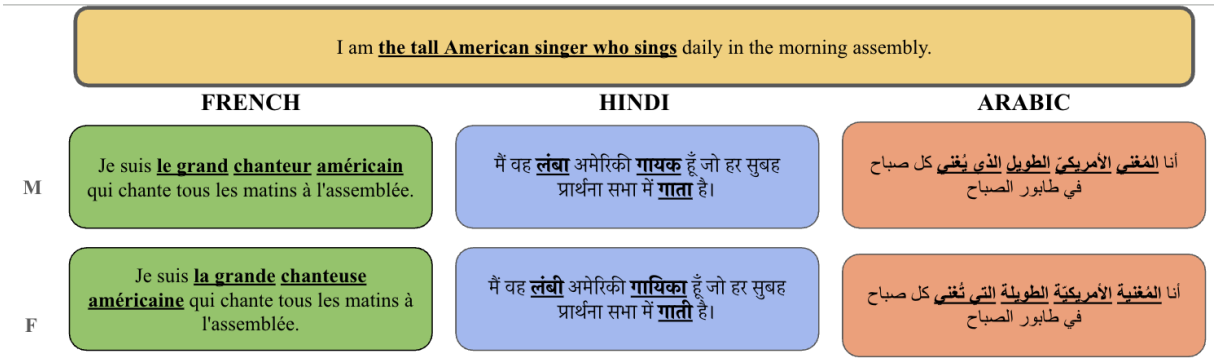
1 Introduction

Multilingual large language models (LLMs) demonstrate strong performance across tasks such as summarization, translation, and question answering (Goel et al., 2023; Qin et al., 2024; Xu et al., 2025; Huang et al., 2025a; Anand et al., 2023). Benchmark datasets like XTREME (Hu et al., 2020), Global-MMLU (Singh et al., 2024b), MM-Eval (Son et al., 2025), BenchMAX (Huang et al., 2025b), and IndicGenBench (Singh et al., 2024a) have become standard tools for evaluating task-specific performance of multilingual LLM models. However, they have been criticized for issues such as poor translation quality, data contamination, and an overwhelming emphasis on high-level tasks that rely heavily on semantic and lexical cues, conflating linguistic competence with broader semantics, making it difficult to isolate fine-grained weaknesses, particularly in morphologically rich or cross-cultural contexts (Wu et al., 2025).

As LLMs are being increasing deployed across diverse linguistic settings, it becomes essential to evaluate their ability to apply morphological rules in a grammatically coherent manner (Piergentili et al., 2024; Savoldi et al., 2025). This is especially critical for languages like French, Arabic, and Hindi, which feature rich grammatical gender constructs, where gender affects verb agreement, pronouns, adjectives, and even word order. For instance, first-person sentences in these languages often contain gendered verbs or adjective forms, even when the subject is implicit (Fig 1). Gender marking in such cases is morphologically subtle yet semantically significant (Gonen et al., 2019). Accurate modeling of gender morphology is thus crucial not only for inclusive applications like conversational agents and machine translation, but also for probing how gender bias manifests in LLMs across gendered language structures (Sitaram et al., 2025; Zhao et al., 2024; Pikuliak et al., 2024). Despite the linguistic and practical importance of gender morphology, there is currently no benchmark that directly evaluates multilingual LLMs on their ability to reason over and apply gender-specific grammatical rules in syntactically rich constructions. Existing work (Joshi et al., 2024; Tang et al., 2025; Sant et al., 2024) has primarily tested morphological competence through tokenization or masked word prediction, but falls short of assessing whether models can generate coherent, grammatical sentences conditioned on gender. To address this gap, we introduce MORPHOGEN a morphologically grounded benchmark dataset covering French, Arabic, and

[*]Authors contributed equally
[†]Corresponding author

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2289–2313



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	अमेरिकी (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

Hindi designed to evaluate gender-conditioned morphological reasoning of LLMs in first-person contexts.

On this benchmark, **we define the GENFORM task as:** given a sentence and the speaker's gender, the model must rewrite the sentence in the opposite gender while preserving grammatical correctness and meaning. To construct linguistically challenging instances, we systematically exploit the rich morphological rules and gender-marking strategies in each language. This requires models to go beyond surface-level transformations and engage in compositional reasoning over linguistic structure. We evaluate *15 widely-used open- and closed-source* multilingual LLMs on this task, spanning model sizes from under *4 billion* to *70 billion parameters*.

Our key contributions are as follows: **(1)** We present a new benchmark and dataset covering three typologically diverse, grammatically gendered languages: French, Arabic, and Hindi, alongside a parallel English corpus for each sentence (Section 3). This setup enables evaluation on our proposed task as well as on related NLP tasks such as machine translation and gender bias analysis. To the best of our knowledge, this is the first and most systematically constructed morphology-focused benchmark for these languages, which we plan to publicly release upon acceptance; **(2)** We introduce novel evaluation metrics to assess the accuracy of gender transformations. These are also applicable to downstream tasks such as translation and gender bias detection in natural language generation (Section 4.2); and **(3)** We benchmark a range of multilingual LLMs on the GENFORM task, providing insights into their ability to model and reason about gendered morphological structures.

2 Related Work

2.1 Existing Benchmarks on Multilingual LLMs

Recent advancements in multilingual LLM evaluation have produced several broad-coverage benchmarks. **XTREME** (Hu et al., 2020) 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** (Son et al., 2025) 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** (Luo et al., 2025) expanded coverage to hundreds of languages across seven NLP tasks, while **mHumanEval** (Raihan et al., 2024) addressed code generation in 200+ languages via machine-translated prompts. Domain-specific ef-

forts like **MuST-SHE** (Bentivogli et al., 2020). and **WinoMT** (Stanovsky et al., 2019) pioneered gender-disambiguated MT datasets for Romance languages, though their narrow scope (1k examples per language) limited utility for LLM evaluation.

2.2 Evaluating Gendered Languages in Multilingual LLMs and NLP Systems

Grammatically gendered languages such as French, Arabic, and Hindi present unique evaluation challenges due to their rich morphological systems. Prior work has shown that large language models (LLMs) often struggle with correctly realizing gender agreement across these languages. For instance, in Hindi, models exhibit errors in gender-inflected verb conjugations and occupational noun morphology (Hada et al., 2024). In Arabic, evaluations reveal gaps in handling gender agreement across dialectal variations (Rhel and Roussinov, 2025), while studies in French demonstrate a tendency for models to default to masculine forms despite contextual cues (Rescigno et al., 2020).

Despite these findings, recent work (Mihaylov and Shtedritski, 2024) highlights that existing benchmarks do not systematically evaluate the application of gender morphology rules across diverse linguistic typologies. This limitation motivates our work. In contrast to prior datasets like Holistic Bias (Smith et al., 2022), which focus on English descriptors of gender and identity, our dataset directly targets the morphological realization of gender in multilingual, grammatically gendered settings.

Complementary lines of research further examine how cultural and speech-related factors influence bias in LLMs and NLP systems, underscoring the importance of addressing gender bias through multiple perspectives and modalities (Goel et al., 2024a; Li et al., 2025; Goel et al., 2024b; Hira et al., 2024).

3 Dataset

In this section, we introduce the MORPHOGEN dataset. We first describe the dataset, the reason for choosing the specific languages, inform what we mean by gendered terms, and explain the rules. Next, we explain the dataset construction process, compare it to existing parallel corpora, and explain the GENFORM task formulation.

3.1 Dataset Description and Statistics

Our proposed dataset, MORPHOGEN, covers three grammatically gendered languages: French, Arabic, and Hindi. For each language, we construct a corpus of sentence pairs. Each sentence pair exists with the first person speaker as masculine and as feminine, along with a parallel English version. Thus, for each sentence, we have its *gender counterfactual* as a ground truth, which is used to define a **Gendered Term**. In other words, gendered terms refer to words that differ between a source sentence and its gender counterfactual.

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

*computed per sentence

Table 1: Statistics for MORPHOGEN dataset

As shown in Table-1, the dataset includes 9,999 French, 2719 Arabic, and 7,610 Hindi sentence pairs. Figure 2 illustrates the distribution of gendered terms per sentence pairs, with some containing up to seven gendered elements, highlighting the morphological complexity of our task.

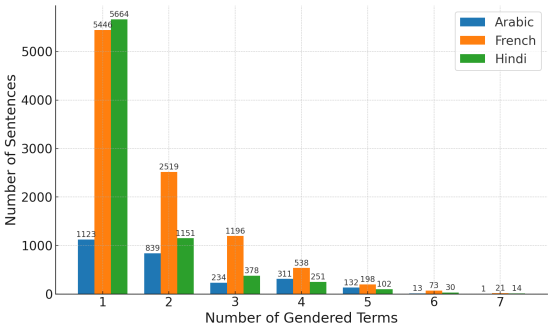


Figure 2: Gendered Terms Distribution in MORPHOGEN

3.2 Task Formulation

For the proposed GENFORM task on MORPHOGEN, we prompt a multilingual LLM with a first-person sentence to rewrite the sentence in the opposite gender, i.e., from masculine to feminine or vice versa, based on the original speaker's gender. The model must correctly apply language-specific morphological rules while preserving the sentence's meaning, fluency, and syntactic structure.

3.3 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

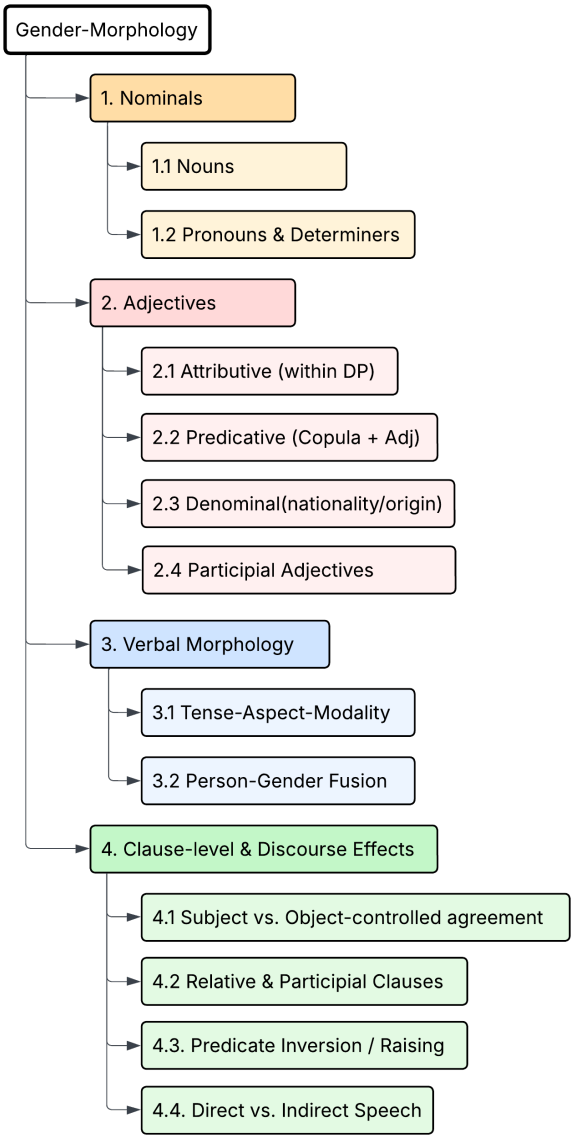


Figure 3: General morphological rules for grammatically gendered languages

assignment strategies of semantic and morphological nature, offering a diverse testbed for evaluating morphological behavior in multilingual LLMs across the selected languages. 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.

3.4 Construction of Morphological Rules

To evaluate the performance of multilingual models on gender transformation in first-person contexts, we constructed a set of language-specific morphological rules grounded in linguistic theory (shown in Figure 4 of Appendix). These rules are inspired by a general taxonomy of gender morphology across grammatically gendered languages (Figure 3) and are illustrated with concrete examples (Table 3 of Appendix). We present an overview of our motivation behind constructing these rules as:

(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 (Lee et al., 2024).

We provide detailed rules with examples for each language in the following tables in the Appendix: French (Tables 4, 5), Arabic (Table 6) and Hindi (Tables 7, 8).

3.5 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. For each language, 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. As each language has its own gender-marking system and grammatical structures, we created language-specific templates (e.g., 'I am a ⟨occupation⟩') and independently generated English source sentences for each language (i.e., the English inputs are not shared across languages), ensuring structural consistency and systematic coverage across cases. 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 (Hurst et al., 2024).These English sentences were translated into Hindi (using IndicTrans2 and GPT-4o-mini) (Gala et al., 2023; Hurst et al., 2024), Arabic (Grok-3)[1], and French (NLLB-200) (Team et al., 2022).

The dataset was refined by multiple bilingual annotators proficient in English and their respective target languages. Each annotator was randomly assigned a subset of the data and instructed to follow the refinement guidelines provided in the appendix, discarding any sentences that did not comply. Subsequently, each sentence was manually corrected into both masculine and feminine forms by the annotators, strictly adhering to the correction guidelines. For detailed annotator instructions, please refer to Appendix B.

Finally, the validity of the dataset was verified by cross-validation among annotators. Every sentence pair was independently reviewed by two annotators. Two evaluation scores were used for this process: the Data Validation Score, which measures the overall proportion of valid samples, and the Inter-Annotator Agreement Score, which reports the fraction of entries where both annotators agreed on the validity judgment. The detailed validation procedure is provided in the Appendix B. Across all three languages, the average Data Validation Score and Inter-Annotator Agreement Score were 0.9705 and 0.9495, respectively. A total of eight annotators in the age group 18–21 participated in this process. Language-wise annotation details are presented in Table 9 in the Appendix. The resulting parallel gender-specific annotations form a high-quality gold-standard set for evaluating the model's sensitivity to morphosyntactic gender variation.

3.6 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 (Koehn, 2005). 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 (Stanovsky et al., 2019). **(2) MT-GenEval** improves diversity and realism for English–Hindi but lacks first-person sentences and speaker-gender labels (Currey et al., 2022). **(3) MuST-SHE** offers speaker annotations and first-person content, but is not publicly available (Bentivogli et al., 2020). **(4) mGENTE** supports gender-neutral generation across languages (Savoldi et al., 2025), but lacks speaker-grounded, first-person constructions. **(5) Arabic Parallel Gender Corpus 2.0** provides first- and second-person gendered sentence pairs from English–Arabic OpenSubtitles (Lison and Tiedemann, 2016; Al Khalifa et al., 2022), but its coverage is limited to a few recurring gender-marking rules.

Our dataset captures a broader range of morphosyntactic phenomena, offering a stronger benchmark across Arabic, Hindi, and French. To the best of our knowledge, no existing dataset:

1. Provides male and female translations for every

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

sentence.
2. Aligns examples with grammatical triggers for gender inflection.
3. Ensures balanced ground truth for both genders.
4. Covers the full spectrum of gender-marking phenomena.

Mining real transcripts is inefficient because most sentences are gender-neutral and only a 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.

4 Experimental Setup

4.1 Models Benchmarked

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 (Grattafiori et al., 2024)
- **Qwen**: Qwen3-4B, Qwen3-8B, Qwen3-14B, Qwen3-27B (Yang et al., 2025)
- **Gemma**: Gemma2-2B, Gemma2-9B, Gemma3-4B, Gemma3-12B, Gemma3-27B (Team et al., 2024, 2025)
- **Phi**: Phi4-14B (Abdin et al., 2024)

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.

4.2 Evaluation Metrics

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 3. 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. SGA captures sentence-level precision in handling gendered terms, ensuring correctness at a fine-grained unit of evaluation. 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 3.2, 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 number 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. This metric is particularly useful for evaluating cases of gender interference, where the model incorrectly alters the gender of words associated with entities other than the explicit speaker. In such cases, GIoU penalizes the model for transforming non-gendered terms, thereby ensuring that only

valid gender-specific modifications are rewarded. 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 number of correctly generated gendered terms across the entire test set to the total number of reference gendered terms in the corpus. This provides a holistic measure of overall transformation quality at an *n*-gram level:

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

Unlike SGA, which evaluates correctness at the sentence level, CGA extends evaluation to the word level across the entire corpus, which makes it especially effective for longer or more complex sentences with multiple gendered terms.

5 Results and Discussion

We evaluated 15 widely-used open-source and closed-source multilingual LLMs on the MORPHOGEN benchmark across French, Arabic, and Hindi, using the metrics defined in Section 4.2. The consolidated results are presented in Table 2, with detailed language-specific results provided in the Appendix: French in Table 10, Arabic in Table 11, and Hindi in Table 12. We analyse the variations in performance across different model families and sizes, and discuss the implications for gender bias in these models.

5.1 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.

5.2 Masculine Bias in French and Arabic

Gender bias varied notably across languages, as seen in the △SGA scores (Figure 5 in Appendix). In Hindi, bias was generally low but occasionally skewed toward feminine forms, with models like Gemma3-4B showing an △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 △SGA of 15.15% due to consistent defaulting to masculine forms. Arabic showed persistent masculine bias, especially in plural constructions, with Qwen3-32B recording an △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 LLMs and underscore the need for targeted debiasing in morphologically rich languages.

5.3 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.

5.4 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 the 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. These patterns are exemplified through results for the LLaMA family of models on multi-entity cases across languages. Illustrative cases for French, Arabic, and Hindi are presented in Figures 7, 8, and 9

Model	French			Arabic			Hindi		
	GIoU ↑	△SGA ↓	CGA ↑	GIoU ↑	△SGA ↓	CGA ↑	GIoU ↑	△SGA ↓	CGA ↑
QWEN2.5-0.5B	5.47	4.55	4.16	4.14	8.49	4.59	0.35	0.63	0.21
GEMMA2-2B	39.73	-5.14	37.54	14.73	-0.81	14.10	71.41	7.35	65.41
LLAMA-3.2-3B	54.49	11.42	53.48	18.31	-27.61	17.75	48.54	-64.65	49.72
GEMMA3-4B	52.70	-14.16	51.60	45.68	-8.20	48.93	67.50	-14.32	64.58
QWEN3-4B	58.64	7.25	53.20	34.34	-0.90	35.97	62.84	3.33	68.51
LLAMA-3.1-8B	67.89	3.69	81.76	43.51	0.96	45.51	83.12	-0.43	89.21
QWEN3-8B	71.66	4.86	69.91	45.89	5.03	51.01	80.96	2.16	87.82
GEMMA2-9B	60.52	1.26	55.56	46.45	2.55	45.26	85.47	-7.34	84.39
GEMMA3-12B	64.27	-0.58	74.26	62.76	2.50	65.52	79.91	-8.93	80.99
PHI4-14B	79.84	1.17	87.70	57.08	6.58	66.15	82.77	1.38	95.10
QWEN3-14B	74.22	14.23	73.91	51.83	9.45	56.08	80.68	7.81	85.80
GEMMA3-27B	71.89	7.53	79.63	70.33	-0.83	74.74	77.97	-7.61	82.56
QWEN3-32B	76.28	10.10	74.74	50.69	11.94	53.00	83.21	5.14	90.38
LLAMA-3.3-70B	76.68	15.15	76.08	59.16	7.50	64.37	93.33	3.67	91.40
GPT-4O-MINI	86.43	-1.11	90.27	71.02	-10.61	80.27	88.81	0.62	93.36

Table 2: Cross-lingual comparison of Gender IoU (GIoU), Sentence-level Gender Accuracy Gap (△SGA), and Corpus-level Gender Accuracy (CGA) across 15 multilingual LLMs for French, Arabic, and Hindi. Higher GIoU and CGA indicate better gender understanding, while lower △SGA indicates reduced bias.

in the Appendix, respectively. 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.

5.5 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 6a 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.

5.6 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 6b 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.

5.7 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 6c 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.

6 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.

7 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. Specifically, we use Modern Standard Hindi, Standard Metropolitan French, and Modern Standard Arabic. 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.

8 Ethical Considerations

While MORPHOGEN aims to advance fairness and inclusivity by providing a gender-focused benchmark for morphologically rich languages (French, Arabic, and Hindi), we recognize several ethical considerations regarding its development and application.

First, our task formulation currently relies on the binary (masculine and feminine) grammatical categories inherent to these languages, which does not encompass the full spectrum of gender identities. We plan to explore non-binary expansions in future iterations, guided by linguistic feasibility and community consultation. Additionally, to avoid reinforcing cultural or occupational stereotypes, we carefully curated prompts to actively challenge male-default biases (e.g., explicitly using feminine forms for roles like "doctor" or "leader").

Second, we acknowledge the general risk that improved grammatical coherence could be misused to generate harmful text. To mitigate this dual-use concern, MORPHOGEN relies exclusively on strictly synthetic prompts and neutral scenarios.

Regarding data creation, annotations were completed by undergraduate students (aged 18–21) who were fairly compensated and certified for their contributions. We prioritized annotator well-being by ensuring all tasks were completely free of sensitive, offensive, or personally identifiable content.

Finally, MORPHOGEN is released under a CC BY-NC 4.0 license for research and non-commercial use, with the intent of helping the community build more equitable and linguistically inclusive NLP systems.

9 Acknowledgments

We would like to acknowledge the Infosys Center for Artificial Intelligence (CAI) and IIIT-Delhi for their support during this research. We are also deeply grateful to Jagjot Singh, Akshit K Bansal, and Ankit Agarwal for their dedicated assistance in the annotation and creation of MORPHOGEN.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero

Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Hend S. Al Khalifa, Abdulmohsen Al-Thubaity, Nora Al-Twairesh, Abdulrahman Alqahtani, and Abdullah Bahanshal. 2022. The arabic parallel gender corpus 2.0: Extensions and analyses. In *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC)*.

Avinash Anand, Arnav Goel, Medha Hira, Snehal Buldeo, Jatin Kumar, Astha Verma, Rushali Gupta, and Rajiv Ratn Shah. 2023. Sciphyrag - retrieval augmentation to improve llms on physics q & a. In *Big Data and Artificial Intelligence: 11th International Conference, BDA 2023, Delhi, India, December 7–9, 2023, Proceedings*, page 50–63, Berlin, Heidelberg. Springer-Verlag.

Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia Antonino Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. Gender in danger? evaluating speech translation technology on the must-she corpus. *Preprint*, arXiv:2006.05754.

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

Jay Gala, Pranjal A Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, and 1 others. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*.

Arnav Goel, Medha Hira, Avinash Anand, Siddhesh Bangar, and Rajiv Ratn Shah. 2023. Advancements in scientific controllable text generation methods. *Preprint*, arXiv:2307.05538.

Arnav Goel, Medha Hira, and Anubha Gupta. 2024a. Exploring multilingual unseen speaker emotion recognition: Leveraging co-attention cues in multitask learning. *Preprint*, arXiv:2406.08931.

Arnav Goel, Medha Hira, and Anubha Gupta. 2024b. Multilingual prosody transfer: Comparing supervised & transfer learning. *Preprint*, arXiv:2406.00022.

Hila Gonen, Yova Kementchedjhieva, and Yoav Goldberg. 2019. How does grammatical gender affect noun representations in gender-marking languages? *Preprint*, arXiv:1910.14161.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Rishav Hada, Safiya Husain, Varun Gumma, Harshita Diddee, Aditya Yadavalli, Agrima Seth, Nidhi Kulkarni, Ujwal Gadiraju, Aditya Vashistha, Vivek Seshadri, and Kalika Bali. 2024. Akal badi ya bias: An exploratory study of gender bias in hindi language technology. *Preprint*, arXiv:2405.06346.

Medha Hira, Arnav Goel, and Anubha Gupta. 2024. Crossvoice: Crosslingual prosody preserving cascade-s2ST using transfer learning. In *The Second Tiny Papers Track at ICLR 2024*.

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

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. 2025a. A survey on large language models with multilingualism: Recent advances and new frontiers. *Preprint*, arXiv:2405.10936.

Xu Huang, Wenhao Zhu, Hanxu Hu, Conghui He, Lei Li, Shujian Huang, and Fei Yuan. 2025b. Benchmax: A comprehensive multilingual evaluation suite for large language models. *Preprint*, arXiv:2502.07346.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ishika Joshi, Ishita Gupta, Adrita Dey, and Tapan Parikh. 2024. 'since lawyers are males..': Examining implicit gender bias in hindi language generation by llms. *arXiv preprint arXiv:2409.13484*.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *ACL Anthology*, pages 79–86.

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

Huihan Li, Arnav Goel, Keyu He, and Xiang Ren. 2025. Attributing culture-conditioned generations to pretraining corpora. *Preprint*, arXiv:2412.20760.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA).

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. 2025. Gloteval: A test suite for massively multilingual evaluation of large language models. *Preprint*, arXiv:2504.04155.

Viktor Mihaylov and Aleksandar Shtedritski. 2024. What an elegant bridge: Multilingual llms are biased similarly in different languages. *Preprint*, arXiv:2407.09704.

Andrea Piergentili, Beatrice Savoldi, Matteo Negri, and Luisa Bentivogli. 2024. Enhancing gender-inclusive machine translation with neomorphemes and large language models. *Preprint*, arXiv:2405.08477.

Matúš Pikuliak, Andrea Hrckova, Stefan Oresko, and Marián Šimko. 2024. Women are beautiful, men are leaders: Gender stereotypes in machine translation and language modeling. *Preprint*, arXiv:2311.18711.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2024. Multilingual large language model: A survey of resources, taxonomy and frontiers. *Preprint*, arXiv:2404.04925.

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

Argentina Anna Rescigno, Eva Vanmassenhove, Johanna Monti, and Andy Way. 2020. 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. 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.

Haneh Rhel and Dmitri Roussinov. 2025. Large language models and arabic content: A review. *Preprint*, arXiv:2505.08004.

Aleix Sant, Carlos Escolano, Audrey Mash, Francesca De Luca Fornaciari, and Maite Melero. 2024. The power of prompts: Evaluating and mitigating gender bias in mt with llms. *Preprint*, arXiv:2407.18786.

Beatrice Savoldi, Eleonora Cupin, Manjinder Thind, Anne Lauscher, Andrea Piergentili, Matteo Negri, and Luisa Bentivogli. 2025. mgente: A multilingual resource for gender-neutral language and translation. *Preprint*, arXiv:2501.09409.

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024a. IndicGenBench: A multilingual benchmark to evaluate generation capabilities of LLMs on Indic languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11047–11073, Bangkok, Thailand. Association for Computational Linguistics.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, and 1 others. 2024b. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *arXiv preprint arXiv:2412.03304*.

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

Eric Michael Smith, Isar Nejadgholi, Ahmad Beirami, and Byron C. Wallace. 2022. "i'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2433–2448, Dublin, Ireland. Association for Computational Linguistics.

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. 2025. Mm-eval: A multilingual meta-evaluation benchmark for llm-as-a-judge and reward models. *Preprint*, arXiv:2410.17578.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Kunsheng Tang, Wenbo Zhou, Jie Zhang, Aishan Liu, Gelei Deng, Shuai Li, Peigui Qi, Weiming Zhang, Tianwei Zhang, and Nenghai Yu. 2025. Gendercare: A comprehensive framework for assessing and reducing gender bias in large language models. *Preprint*, arXiv:2408.12494.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

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, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Minghao Wu, Weixuan Wang, Sinuo Liu, Huifeng Yin, Xintong Wang, Yu Zhao, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. The bitter lesson learned from 2,000+ multilingual benchmarks. *Preprint*, arXiv:2504.15521.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Jinman Zhao, Yitian Ding, Chen Jia, Yining Wang, and Zifan Qian. 2024. Gender bias in large language models across multiple languages. *arXiv preprint arXiv:2403.00277*.

A Gender-Morphology

Different languages express grammatical gender through distinct morphological patterns. An overview of these patterns is shown in Figure 3, with illustrative examples in Table 3. These patterns motivate our focus on three gendered languages: French, Arabic, and Hindi.

To support comparisons made in the results section, we define morphological complexity in terms of (i) the number of agreement targets (e.g., verbs, adjectives, determiners), (ii) the regularity vs. irregularity of gender marking, and (iii) the extent to which gender realization depends on syntactic context (e.g., tense, clause structure, or discourse configuration). Under this definition, French and Arabic both exhibit high morphological complexity, but for different reasons. French shows irregular and context-dependent agreement (e.g., gender agreement in past participles but not in present tense), while Arabic exhibits a more systematic yet nuanced morphology with pervasive agreement across verbs, adjectives, pronouns, and constructions such as relative clauses, conditionals, and multi-entity contexts. In contrast, Hindi follows a comparatively more regular and semantically grounded system, with fewer context-dependent variations.

For each of these languages, we provide example snippets along with the corresponding morphological rules in Tables 4, 5, 6, 7, and 8.

B Annotator Guidelines

Guidelines for Dataset Refinement

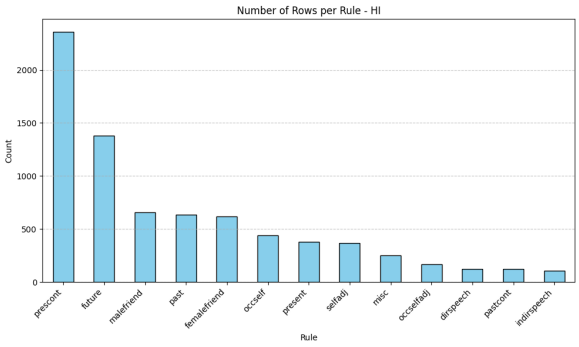
- **Off-Limits Language:** No sentence may contain profanity, hate speech, slurs, or any other abusive or objectionable content. Annotators must ensure compliance with content-policy restrictions.
- **Naturalness:** Sentences should reflect standard conversational phrasing that a fluent speaker would naturally use, avoiding stilted or machine-generated constructions.
- **Uniqueness:** Identical or trivially paraphrased sentences are to be rejected and regenerated.
- **Template Fidelity:** Sentences must follow the syntactic template exactly, without missing slots, extra words, or rearrangements.
- **Domain Coverage:** For every template, sentences must span all conversational domains specified (e.g., academic, healthcare, legal), ensuring diversity.
- **Gender Specificity:** Each English sentence must be designed such that its translations differ between masculine and feminine forms in the target language.

Guidelines for Dataset Correction

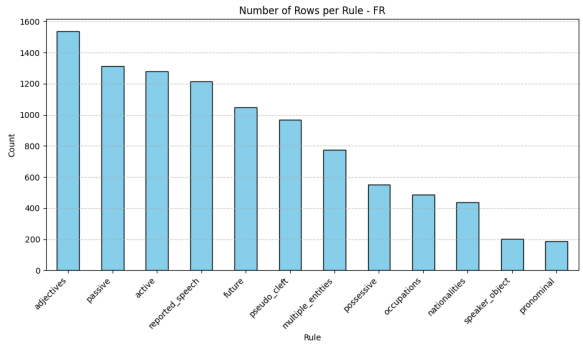
- **Fidelity & Fluency:** Translations must preserve meaning, tone, and register while being grammatically correct and idiomatic in the target language. Annotators should check word choice, tense, punctuation, and readability.
- **Speaker-Gender Agreement:** All gender-dependent morphology tied to the speaker (verbs, adjectives, pronouns, etc.) must appear in the correct masculine form in the "male" version and the correct feminine form in the "female" version.
- **Consistency for Implicit Gender Entities:** Gendered terms referring to non-speaker entities must remain identical across male and female translations. For instance, if *friend* is rendered in masculine form in the male version, it must remain masculine in the female version as well.

Dataset Validation Process

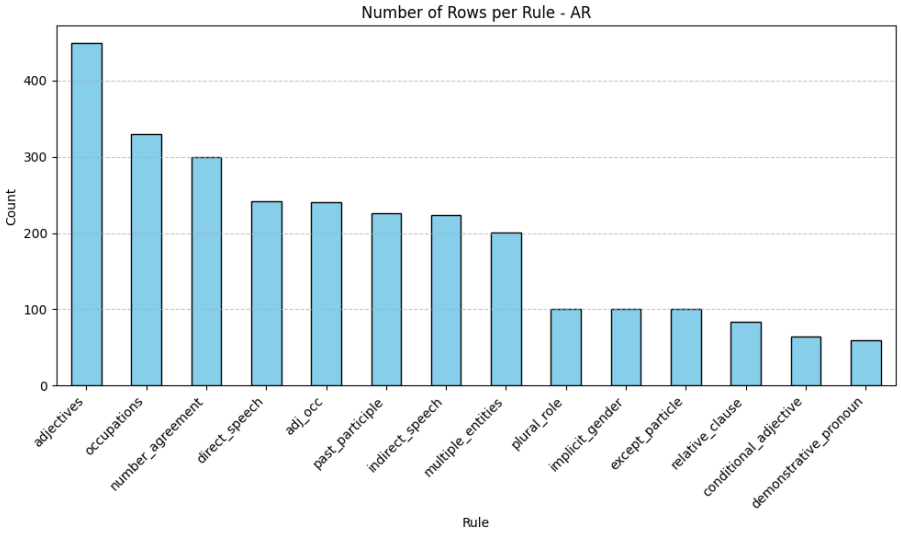
Each data sample was independently assigned a validity score of 1 or 0 by two annotator, indicating full compliance or non-compliance with the annotation guidelines, respectively. The Data Validation



(a) Hindi



(b) French



(c) Arabic

Figure 4: Distribution of Sentence Frequency Per Morphological Rule for Each Language

Score (DVS) and Inter-Annotator Agreement (IAA) were computed as follows:

$$\text{DVS} = \frac{\sum_{i=1}^{N} (s_{i1} + s_{i2})}{2N} \quad (1)$$

$$\text{IAA} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(s_{i1} = s_{i2}\right) \quad (2)$$

where:

- N denotes the total number of sentence pairs in the dataset.
- s_{i1} and s_{i2} represent the binary validity scores (0 or 1) assigned to the i^{th} sentence pair by Annotator 1 and Annotator 2, respectively.
- $\mathbb{I}(\cdot)$ is the indicator function, returning 1 if the condition inside is true and 0 otherwise.

The *Data Validation Score* measures the overall proportion of valid samples across all annotators, while the *Inter-Annotator Agreement* quantifies the fraction of samples for which both annotators assigned identical scores.

C 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`

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 3: 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 4: 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 5: 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 يوميًا. أنا معلمة مخلصة ألهم الطلاب يوميًا.
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 6: 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 7: 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 8: Hindi Gendered Grammar Examples Across Rule Types [2]

Statistics	Arabic	French	Hindi
Total Sentences Generated	5413	16415	10248
Sentences Discarded	2638	6416	2694
Unique Sentences	7610	9999	10248
Data Validation Score	0.9733	0.9651	0.9731
Inter-Annotator Score	0.9526	0.9366	0.9594
Number of Annotators	2	3	3

Table 9: Dataset Statistics Across 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	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 10: Performance metrics of different models on French (% values; GIoU = Gender IoU, CGA = Corpus-Level Gender Accuracy, SGA = Sentence-Level Gender Accuracy, $\triangle$SGA = Accuracy Gap, M = Male, F = Female)

- **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.

D Prompts

D.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
```

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	**-27.61**	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 11: Performance metrics of different models on Arabic (% values; GIoU = Gender IoU, CGA = Corpus-Level Gender Accuracy, SGA = Sentence-Level Gender Accuracy, $\triangle$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	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 12: Performance metrics of different models on Hindi (% values; GIoU = Gender IoU, CGA = Corpus-Level Gender Accuracy, SGA = Sentence-Level Gender Accuracy, $\triangle$SGA = Accuracy Gap, M = Male, F = Female)

`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.

D.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.`

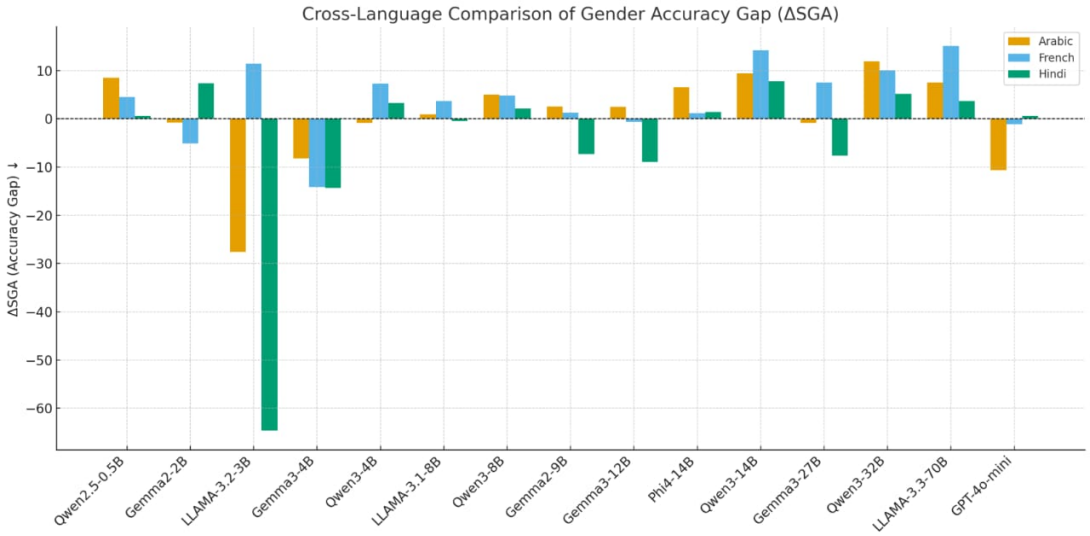


Figure 5: △SGA (Accuracy Gap) across all models and languages (French, Arabic, Hindi) in the MORPHOGEN benchmark. Positive values indicate masculine bias, while negative values indicate feminine bias.

- 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.

E Detailed Results and Error Analysis

This section provides a comprehensive breakdown of model performance on the MORPHOGEN benchmark, combining dataset statistics, rule-level quantitative metrics, and a qualitative error analysis of the LLAMA model family.

E.1 Quantitative Performance and Bias Analysis

Table 9 details the dataset validation statistics, while the aggregate performance metrics across all models for French, Arabic, and Hindi are presented in the respective language tables.

To better understand directional bias, Figure 5 visualizes the Accuracy Gap (ΔSGA) across all models. A clear trend emerges regarding model scale and gender bias. Smaller models exhibit erratic and often extreme bias gaps. For instance, smaller architectures like LLAMA 3.2 3B demonstrate massive fluctuations, including severe feminine bias ($\Delta\text{SGA} < 0$) in Hindi and Arabic, contrasted with masculine bias in French. As model capacity increases (e.g., LLAMA 3.3 70B and GPT-4O-MINI), the ΔSGA converges closer to zero across all languages, indicating a more balanced, unbiased linguistic understanding rather than a reliance on statistical gender defaults.

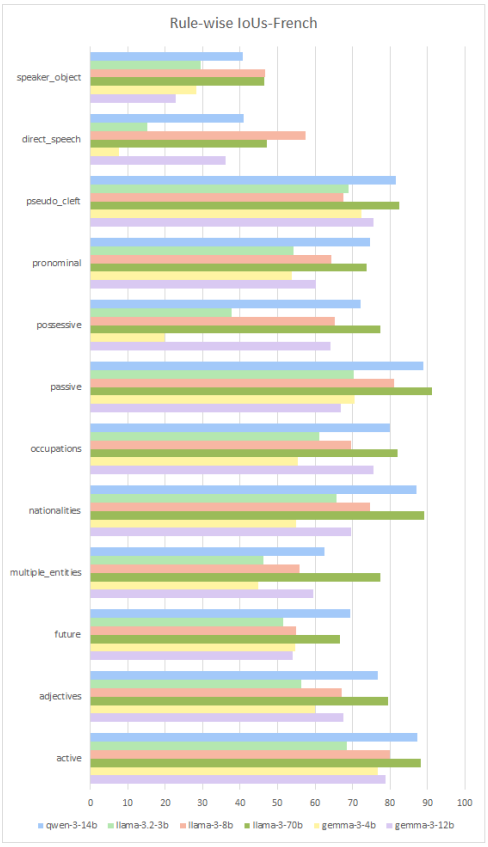
E.2 Rule-Level Morphological Competence

To isolate where models succeed or fail, Figure 6 presents the Gender Intersection over Union (GIoU) metrics broken down by specific grammatical rules for all 3 languages.

Across all three languages, models generally perform well on localized, simple rules (e.g., basic adjectives). However, performance sharply degrades on complex syntactical structures that require long-range dependency tracking, such as indirect speech, pseudo-cleft sentences, and multiple entities. The gap between smaller and larger models is most pronounced in these complex categories, highlighting that parameter scale is crucial not just for vocabulary, but for maintaining morphological consistency across extended contexts.

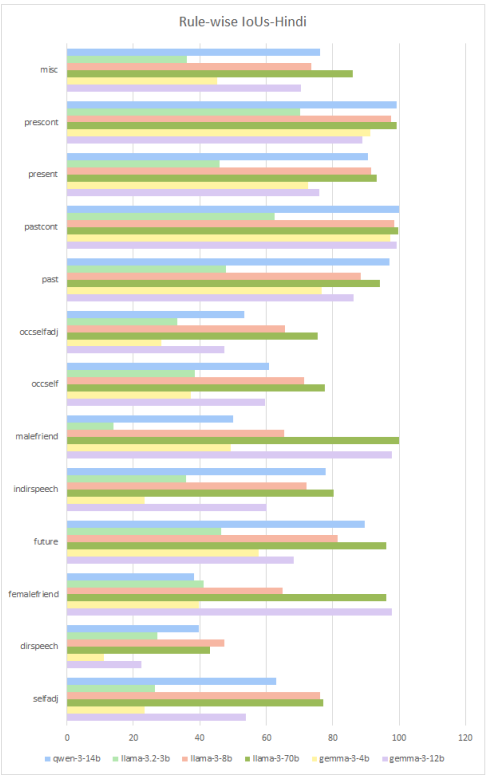
E.3 Qualitative Error Analysis: A LLAMA Case Study

To ground these quantitative findings, Figures 7, 8, and 9 present a qualitative error analysis focus-



(a) French

(b) Arabic



(c) Hindi

Figure 6: Rule-based and model-wise IoU metrics across all three languages.

Source (English): Even though I was understood by my friend, I still felt the need to elaborate.

Reference: Même si j'ai été **comprise** (f) par mon ami (m), j'ai toujours senti le besoin d'élaborer.

Contrastive: Même si j'ai été **compris** (m) par mon ami (m), j'ai toujours senti le besoin d'élaborer.

LLAMA 3.2 · 3B (SGA↓ GIoU↓)

Generated: Même si j'ai été **compris** (m) par mon ami (m), j'ai toujours senti le besoin d'élaborer.

LLAMA 3.1 · 8B (SGA↑ GIoU↓)

Generated: Même si j'ai été **comprise** (f) par mon **amie** (f), j'ai toujours senti le besoin d'élaborer.

LLAMA 3.3 · 70B (SGA↑ GIoU↑)

Generated: Même si j'ai été **comprise** (f) par mon ami (m), j'ai toujours senti le besoin d'élaborer.

Underlined English words are those whose **French** inflection reflects gender of the associated entity. **Green** and **Red** mark the correct and the incorrect gender form for the speaker respectively; **Purple** marks the incorrect form for the *other* entity in the sentence. Markers *(m)* and *(f)* denote masculine and feminine grammatical forms, respectively.

Figure 7: Example of results of LLAMA family of models on multiple entities in French dataset.

Source (English): I am a teacher, my neighbor is a painter, she creates art, and my mentor is a scientist.

Reference: أنا معلمة (f)، جارتي رسامة، هي تخلق فنا (m)، ومرشدي عالم (m).

Contrastive: أنا معلم (m)، جارتي رسامة، هي تخلق فنًا (m)، ومرشدتي عالمة (m).

LLAMA 3.2 · 3B (SGA↓ GIoU↓)

Generated: أنا معلم (m)، جارتي رسامة، هي تخلق فنًا (m)، ومرشدي عالم (m).

LLAMA 3.1 · 8B (SGA↑ GIoU↓)

Generated: أنا معلمة (f)، جارتي رسامة، هي تخلق فنا (f)، ومرشدتي عالمة (f).

LLAMA 3.3 · 70B (SGA↑ GIoU↑)

Generated: أنا معلمة (f)، جارتي رسامة، هي تخلق فنًا (m)، ومرشدي عالم (m).

Underlined English words are those whose **Arabic** inflection reflects gender of the associated entity. **Green** and **Red** mark the correct and the incorrect gender form for the speaker respectively; **Purple** marks the incorrect form for the *other* entity in the sentence. Markers *(m)* and *(f)* denote masculine and feminine grammatical forms, respectively.

Figure 8: Example of results of LLAMA family of models on multiple entities in Arabic dataset.

Source (English): I know her from time to time, my friend serves as a senior administration officer.

Reference: मैं उसे काफ़ी समय से **जानता** (m) हूँ मेरी (f) दोस्त वरिष्ठ प्रशासन अधिकारी का काम करती (f) है।

Contrastive: मैं उसे काफ़ी समय से **जानती** (f) हूँ मेरी (f) दोस्त वरिष्ठ प्रशासन अधिकारी का काम करती (f) है।

LLAMA 3.2 · 3B (SGA↓ GIoU↓)

Generated: मैं उसे काफ़ी समय से **जानती** (f) हूँ मेरी (f) दोस्त वरिष्ठ प्रशासन अधिकारी का काम करती (f) है।

LLAMA 3.1 · 8B (SGA↑ GIoU↓)

Generated: मैं उसे काफ़ी समय से **जानता** (m) हूँ **मेरा** (m) दोस्त वरिष्ठ प्रशासन अधिकारी का काम **करता** (m) है।

LLAMA 3.3 · 70B (SGA↑ GIoU↑)

Generated: मैं उसे काफ़ी समय से **जानता** (m) हूँ मेरी (f) दोस्त वरिष्ठ प्रशासन अधिकारी का काम करती (f) है।

Underlined English words are those whose **Hindi** inflection reflects gender of the associated entity. **Green** and **Red** mark the correct and the incorrect gender form for the speaker respectively; **Purple** marks the incorrect form for the *other* entity in the sentence. Markers *(m)* and *(f)* denote masculine and feminine grammatical forms, respectively.

Figure 9: Example of results of LLAMA family of models on multiple entities in Hindi dataset.

ing on "gender interference" in complex sentences containing multiple entities. We compare the outputs of LLAMA 3.2 3B, LLAMA 3.1 8B, and LLAMA 3.3 70B to illustrate the evolution of morphological control. Further structural examples of grammatical rules for each language can be found in Tables 5 through 8.

LLaMA 3.2 3B. The smallest model consistently struggles to correctly apply gendered inflections, often defaulting to standard forms regardless of the intended speaker gender. This behavior is evident across all three languages, where the model fails to align verbs, adjectives, or participles with the correct grammatical gender. As seen in the qualitative examples, these persistent errors in speaker gender realization heavily penalize its overall SGA and GIoU scores, and explain the extreme ΔSGA variations observed in Figure 5.

LLaMA 3.1 8B. The 8B model demonstrates a clear improvement in capturing basic gender morphology, particularly in correctly inflecting verbs and adjectives immediately adjacent to the speaker. However, it suffers from overgeneralization, leading to severe *gender interference*. As illustrated in Figures 7 through 9, while the primary speaker's gender is correctly realized, the model incorrectly alters the grammatical gender of *other* entities in the sentence to match the speaker. This indicates a partial understanding of agreement rules but insufficient syntactic control over entity-specific boundaries. Consequently, while its SGA scores improve relative to the 3B model, its GIoU remains suppressed due to these collateral inflection errors.

LLaMA 3.3 70B. The largest model exhibits robust performance across all examined cases, correctly applying gender transformations while strictly preserving agreement boundaries between multiple entities. It maintains a sharp distinction between speaker-specific and non-speaker-specific gender marking, correctly inflecting only the relevant tokens without "bleeding" the gender onto adjacent nouns. This results in outputs that closely match the reference sentences both structurally and morphologically. Correspondingly, the model achieves the highest SGA and GIoU scores (as seen in the rule-wise plots (Fig 6) and minimal bias gaps, reflecting highly accurate gender realization across complex grammatical scenarios.