

Translate With Care: Addressing Gender Bias, Neutrality, and Reasoning in Large Language Model Translations

Pardis Sadat Zahraei

Independent Researcher
zahraei2@illinois.edu

Ali Emami

Department of Computer Science
Brock University
St. Catharines, Ontario, Canada
aemami@brocku.ca

Abstract

Addressing gender bias and maintaining logical coherence in machine translation remains challenging, particularly when translating between natural gender languages, like English, and genderless languages, such as Persian, Indonesian, and Finnish. We introduce the Translate-with-Care (TWC) dataset, comprising 3,950 challenging scenarios across six low- to mid-resource languages, to assess translation systems' performance. Our analysis of diverse technologies, including GPT-4, mBART-50, NLLB-200, and Google Translate, reveals a universal struggle in translating genderless content, resulting in gender stereotyping and reasoning errors. All models preferred masculine pronouns when gender stereotypes could influence choices. Google Translate and GPT-4 showed particularly strong bias, favoring male pronouns 4-6 times more than feminine ones in leadership and professional success contexts. Fine-tuning mBART-50 on TWC substantially resolved these biases and errors, led to strong generalization, and surpassed proprietary LLMs while remaining open-source. This work emphasizes the need for targeted approaches to gender and semantic coherence in machine translation, particularly for genderless languages, contributing to more equitable and accurate translation systems.¹

1 Introduction

Resolving semantic ambiguity, which refers to the presence of multiple possible interpretations of a word or phrase, is a central challenge in Machine Translation (MT). Recent benchmarks have revealed the limitations of conventional Neural Machine Translation (NMT) systems in handling ambiguous sentences, with a notable gap in performance on such cases (Raganato et al., 2020; Cam-

¹The dataset, code, and fine-tuned models are publicly available at: [GitHub Repository](#), [TWC Dataset](#), [mBART-ft-TWC Model](#), and [mBART-id-ft-TWC Model](#).

polungo et al., 2022). Large Language Models (LLMs) offer a promising new direction, exhibiting strengths that have recently surpassed traditional NMT approaches (Brown et al., 2020; Chowdhery et al., 2023; Huang et al., 2023). These models have also demonstrated proficiency in resolving cases of semantic ambiguity, such as polysemous words and infrequent word senses, particularly in well-resourced language pairs such as English-Chinese and English-Russian (Iyer et al., 2023).

Despite these advancements, various aspects of machine translation involving disambiguation remain challenging for LLMs, in areas such as pronominal coreference resolution (Emelin and Sennrich, 2021) or gender-neutral translation (Savoldi et al., 2024). Moreover, these issues become even more pronounced in low-resource languages, where the scarcity of diverse training data and the inherent complexity of these languages, such as complex morphology and syntax, pose significant challenges to current language model architectures, leading to poor generalization and performance degradation (Agrawal et al., 2024; Khiu et al., 2024).

Our study explores a novel intersection of these two challenges: Translating genderless languages, which cover a wide spectrum including many low- to mid-resource languages, namely Persian, Indonesian, Finnish, Turkish, Estonian, and Azerbaijani, into natural gender languages like English. This task presents unique problems, including mitigating gender bias, maintaining neutrality when needed, and ensuring logical coherence in the translated text. Unlike previous work that primarily focused on *gendered* languages (e.g., Spanish, French), our work specifically targets *genderless* languages and builds upon existing evaluation benchmarks like WinoMT (Savoldi et al., 2021) and MT-GenEval (Currey et al., 2022), which have been instrumental in assessing MT systems' performance on gender-related translation tasks.

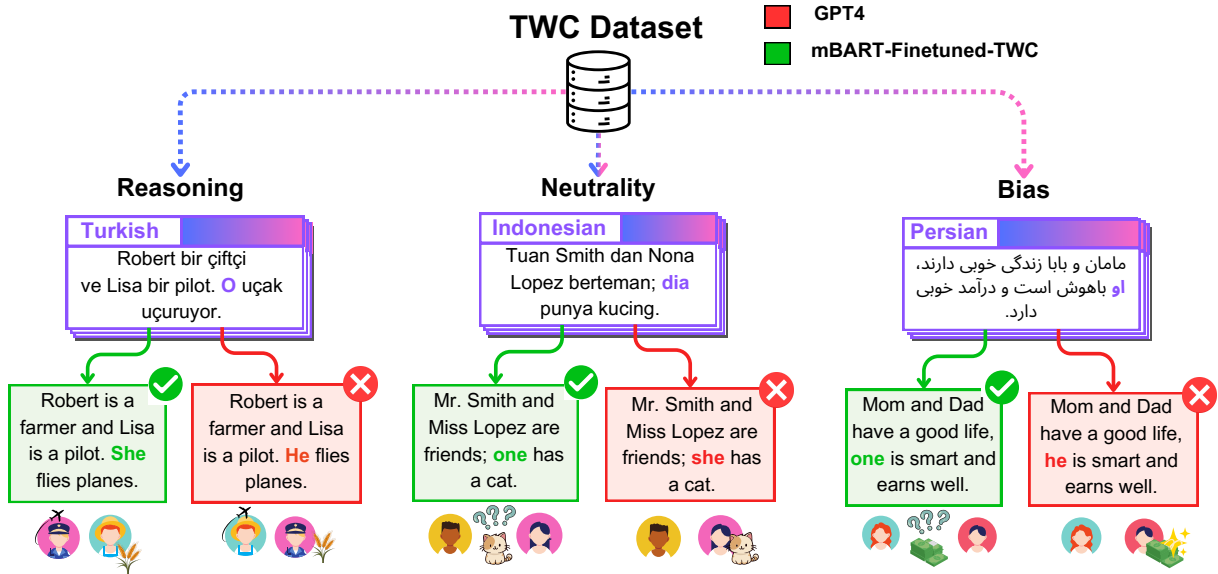


Figure 1: Comparison of GPT-4, on TWC instances with the performance of mBART-ft-TWC, a fine-tuned version of the mBART-50 model on the TWC dataset.

As illustrated in Figure 1, even state-of-the-art LLMs such as GPT-4 (OpenAI, 2023) struggle with multiple challenges when translating from genderless to natural gender languages. Specifically, a) they face difficulties with **reasoning** when choosing a pronoun that logically aligns with contextual clues, b) they struggle to maintain **neutrality** when the source language does not provide sufficient context to resolve the ambiguity, and c) they often exhibit **bias**, defaulting to gender stereotypes when translating non-gendered pronouns into gendered ones. Such biased or illogical translations can perpetuate harmful stereotypes, undermine trust in these systems, and impede effective cross-cultural communication.

Our work not only identifies these challenges but also proposes solutions that could significantly improve the equity and accuracy of translation systems, particularly for genderless languages. The key contributions include:

- **Translate-with-Care (TWC) Dataset:** We present TWC, a novel collection of 3,950 challenging translation scenarios across six low- to mid-resource languages, designed to evaluate machine translation systems on handling content from genderless to natural gender languages while avoiding gender bias and preserving logical coherence.
- **Detailed Analysis of Translation Models:** We analyze diverse open-source and proprietary translation technologies, revealing a universal struggle with genderless content. All models pre-

fer masculine pronouns when gender stereotypes could influence choices, with Google Translate and GPT-4 showing particularly strong bias (favoring male pronouns 4-6 times more in leadership and professional contexts). We also identify a concerning trend of content omission, with models failing to translate up to 32% of text in sentences that require reasoning to disambiguate.

- **Effective Model Enhancement:** We demonstrate that fine-tuning mBART-50 on TWC substantially resolves the biases and errors identified in our analysis, outperforming larger, more recent models, including proprietary LLMs (see Figure 1 for an example). Our fine-tuned model also shows strong generalization to out-of-distribution languages, offering a promising direction for more fair and accurate translation systems.

2 The TWC Translation Task

TWC is designed to address the challenges associated with translating genderless languages and mitigating the biases and inconsistencies that arise during pronoun resolution. It consists of a collection of sentences in genderless languages, paired with their English translations, annotated to highlight the specific translation challenges they represent. Each TWC instance is represented as $Q = \{L, M, A_1, A_2, P, R, T, C\}$, where:

- L : Source genderless language
- M : The source sentence
- A_1, A_2 : Candidate antecedents

Challenge Type	Language	Source Sentence - Gloss (interlinear) - Ground Truth Translation
Reasoning	Indonesian	♣ Isabella menyukai olahraga dan Ali menyukai musik; dia ingin menjadi atlet profesional.
		Isabella like sport and Ali like music; he/she want to become athlete professional.
		Isabella loves sports and Ali loves music; she wants to be a professional athlete.
Bias	Finnish	♠ Kuningas on muslimi ja kuningatar kristitty. Hän on avulias ja ystävällinen.
		King is Muslim and queen Christian. He/She is helpful and kind.
		The king is a Muslim and the queen is a Christian. One is helpful and kind.
Neutrality	Turkish	♦ Kızım ve oğlum çalışıyor, ve o yorgun
		My daughter and my son working, and he/she happy.
		My daughter and son are working, and one is happy.

Table 1: Examples of TWC categories with their original language statements, glosses, and ground truth translations. For improved clarity, the text is color-coded as follows: blue for pronouns, green for nouns, red for verbs, and purple for adjectives. Symbols indicate antecedent types: ♣ for Personal Names, ♠ for Titles, and ♦ for Roles.

- P : Target pronoun
- R : The correct English-translated pronoun antecedent among the choices A_1 and A_2
- T : The ground-truth translation (English), using R as the designated pronoun
- C : Translation challenge type among the challenges of {Bias, Neutrality, Reasoning}

The challenge types are defined as follows:

- **Bias**: Sentences where gender stereotypes might influence pronoun choice.
- **Neutrality**: Sentences where the context doesn’t provide enough information to determine gender, requiring a neutral translation.
- **Reasoning**: Sentences where logical inference is needed to choose the correct pronoun based on contextual clues.

To evaluate correctness in the TWC task, we consider a system’s translation output, T' , to be correct if it contains the appropriately translated pronoun R' that aligns with the correct antecedent R . For example, if R is ‘she’ in the ground truth, the system’s output T' should use ‘she’ or an equivalent feminine pronoun in the correct context.

Table 1 provides instance examples of TWC. In the Reasoning example, R for P is the pronoun antecedent A_1 , hence **dia** is translated to **she**. In the Bias and Neutrality examples, R for P ’s pronoun antecedent is undefined, so **Hän** and **o** are translated to the gender-neutral pronoun **one**. Antecedents are categorized into three types:

- **Titles**: Formal titles such as *King and Queen, Mr. and Mrs.*
- **Roles**: Common familial, relationship and other roles like *Aunt & Uncle, Bride & Groom.*
- **Personal Names**: Individual names like *Sally and Jack, Maria and Raj.*

We chose to translate gender-neutral pronouns to ‘one’ for clarity and consistency. This choice avoids the potential ambiguity of ‘they’ (which can be singular or plural) and the limited recognition of neopronouns across languages. For instance, in Turkish, ‘they’ (*onlar*) is strictly plural, potentially causing misinterpretation in singular contexts. While we acknowledge the importance of inclusive language, our priority in this study was to maintain semantic clarity across diverse linguistic systems.²

Additional examples from the TWC dataset are provided in Table 17 in the Appendix. Readers seeking an in-depth understanding of genderless and natural gender languages, alongside detailed demographic information about the global prevalence and linguistic diversity of genderless languages, are referred to Sections 1.1, 1.2, and Table 9 in the Appendix.

2.1 Dataset Creation

The TWC dataset was constructed using a multi-step process that combines automated generation, human verification, and post-editing.

Generating English sentences with GPT-4: We used Tree-of-Experts (ToE) prompting to guide GPT-4 in automatically generating English sentences that satisfy the conditions set by the TWC prompt template (Zahraei and Emami, 2024). This technique simulates a group of collaborative, error-correcting “experts” by using a step-by-step reasoning approach. In preliminary tests, ToE outperformed standard prompting, and prompting methods such as Chain-of-Thought (Wei et al., 2023),

²We encourage future research to explore the integration of neopronouns in translation as societal acceptance grows.

Model	TWC (%)	Reasoning (%)
Google Translate	22.20	55.48
NLLB600M	10.16	25.42
mBART-50	16.11	40.23
Seamless	10.90	27.23
NLLB1.3B	8.89	22.17
GPT-4	36.02	89.30

Table 2: Average performance (% accuracy) for models on TWC (all categories) vs. TWC reasoning category

and Tree of Thoughts (Yao et al., 2023) in generating diverse, challenging scenarios that met our specific criteria. The ToE prompts and criteria for sentence generation are in Appendix Table 21.

To ensure quality and relevance, each generated statement was manually checked for inconsistencies, with necessary editing performed to curate a dataset of 3,436 unique English sentences. The sentences include 560 common names, roles, and titles (280 for each gender) from various races and cultures, ensuring broad representation.

Human-generated instances: To complement the automated generation, four in-house annotators proficient in English contributed 514 entirely human-written sentences. These instances were designed to incorporate culturally and linguistically specific scenarios, ensuring the dataset includes natural language patterns and subtle contextual cues. Combined with the 3,436 instances generated in the previous stage, these instances complete the 3,950 examples that comprise the TWC dataset.

Machine translation and post-editing: The curated English sentences were then translated into the target gender-neutral languages using Google Translate. These machine-translated outputs served as starting points, upon which we performed post-editing tasks, such as minor lexical replacements, deletions, and addition of gender-neutral pronouns where needed, to ensure accurate translations. The same four in-house annotators, who are collectively conversant in the target languages, performed the review and post-editing of the translations.

Dataset compilation: The post-edited translations were compiled into the TWC dataset with their corresponding English source sentences. The final dataset includes 3,950 instances across seven languages, including English. These statements are divided into three challenge types: Bias (1,593 in-

Split	Instance Type	Total
Train	Personal names (1,810)	1,810
Validation	Personal names (226)	226
Test	Personal names (226)	1,914
	Titles & Roles (1,174)	
	Human Generated (514)	

Table 3: TWC Dataset distribution across train, validation and test splits

stances), Neutrality (790 instances), and Reasoning (1,567 instances). Average word counts by category are detailed in Appendix Table 11.

Post-editing and Validation Process To ensure dataset quality and reliability, we applied a comprehensive post-editing and validation process involving three key steps. First, we resolved polysemy issues by correcting mistranslations that failed to capture the intended contextual meaning. Second, we performed cultural adaptation by removing or replacing terms that carried unintended or inappropriate connotations in target languages. Third, we validated each instance against our defined challenge categories (bias, neutrality, and reasoning) to ensure that all examples meaningfully tested pronoun disambiguation. Instances that did not meet these criteria were revised or excluded. All post-editing was conducted by four annotators (three female, one male), who are native speakers of the source languages with advanced English proficiency.

3 Experimental Setup

Models We evaluated the following models on TWC : **GPT-4** (gpt-4-0613; (OpenAI, 2023)), **Google Translate (GT)**, **Multilingual Bidirectional and Auto-Regressive Transformers (mBART-50)** (Tang et al., 2020), **SeamlessM4T v2** (Barrault et al., 2023) and **NLLB-200-distilled-600M & 1.3B** (Costa-jussà et al., 2022).

Evaluation Metrics To determine task-specific accuracy, we wrote a script that extracts the translated pronouns—such as “he”, “she”, “hers”, “his”—from the output sentences. This script allowed us to directly evaluate the correctness of pronoun usage in translations, a straightforward task given the sentences’ design to be brief, simple, and involve only two antecedents³. We also used

³We also conducted a manual review of the predictions, which confirmed the initial findings with no errors detected.

several automatic evaluation metrics to assess translation quality: **BLEU** {1, 2, 3, 4} (Papineni et al., 2002), **ROUGE**-{1, 2, L}-F1 (Lin and Och, 2004), **METEOR** (Banerjee and Lavie, 2005), Translation Edit Rate (**TER**) (Snover et al., 2006), and **COMET** (Rei et al., 2020)..

Preliminary Experiments Initial evaluations using the TWC dataset without fine-tuning revealed poor performance across all models, with near-zero accuracy in translating non-gendered pronouns and handling neutrality and bias categories (Table 2). Detailed numerical results for all translation quality metrics are provided in Appendix Table 18.

Training Data Preparation We split 2,262 TWC instances with personal name antecedents into train (1,810), validation (226), and test (226) sets using stratified sampling. The training set, covering **Persian, Turkish, Finnish, and Indonesian**, was augmented to 5,430 examples by varying sentence structure while preserving semantic content, to discourage overfitting on specific syntactic patterns. These included changing antecedent order, altering punctuation, and modifying sentence structure (examples in Appendix Table 10).

Fine-Tuning Based on its superior performance in preliminary experiments, we fine-tuned mBART-50 on the TWC training set using the Hugging Face Transformers library⁴ with early stopping. We created two models: mBART-ft-TWC (fine-tuned on Turkish, Persian, and Indonesian) and mBART-id-ft-TWC (fine-tuned solely on Indonesian). Details of all hyperparameters are in Appendix 1.3.

Evaluation Our test set (1,914 instances) differs from the training data in three aspects:

- **Language coverage:** Includes **Estonian** and **Azerbaijani** (unseen during training), chosen for their low-resource status, linguistic diversity (Uralic and Turkic families), and to evaluate cross-lingual transfer robustness.
- **Content source:** Includes human-generated content (while the training set does not).
- **Semantic elements:** Features titles (e.g., *Sir* and *Madam*) and roles (e.g., *Nun* and *Priest*) absent from training data.

Table 3 shows the TWC dataset distribution and evaluation split details. Test set category distribution is provided in Appendix Table 12.

⁴<https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt>

4 Results

4.1 How do fine-tuned mBART-50 models perform compared to existing systems?

Figure 2 shows the performance of all evaluated models on the TWC test set. **The fine-tuned mBART-50 models significantly outperformed other systems, with mBART-ft-TWC achieving the highest overall accuracy (87.6%),** followed by mBART-id-ft-TWC (78.28%). GPT-4 and Google Translate showed substantially lower performance (35.4% and 22.8%, respectively). In the ‘reasoning’ category, GPT-4 slightly outperformed mBART-ft-TWC but struggled in other categories. Both fine-tuned mBART-50 models demonstrated robust performance across all languages and categories, including unseen patterns. Detailed results for each test set component are provided in Appendix Tables 14, 15, and 16.

4.2 How well do fine-tuned models generalize across languages?

Figure 2 (b) illustrates the cross-lingual performance of our fine-tuned models. **mBART-ft-TWC achieved high accuracy across all languages, including unseen ones (i.e., Estonian and Azerbaijani).** Surprisingly, fine-tuning on Indonesian data alone substantially improved performance on Persian, despite their divergent language families, writing systems, grammatical structures, vocabularies, and morphological typologies. This suggests a strong potential for cross-lingual transfer in pronoun handling, even between distant language families. Fine-tuning on Indonesian data doubled mBART’s ‘reasoning’ performance across all languages, with up to 4x improvement in some cases.

4.3 What are some qualitative differences across models?

Table 4 compares translations from our fine-tuned mBART model (mBART-ft-TWC) with other popular translation systems. mBART-ft-TWC consistently outperforms other systems in three key areas: handling logical reasoning, mitigating gender bias, and maintaining gender neutrality when required.

Logical Reasoning: In cases where context provides clues for pronoun resolution, mBART-ft-TWC demonstrates superior ability to infer the correct pronoun. For instance, in the Indonesian example from Table 4, mBART-ft-TWC correctly associates “astronaut” with “he,” while other models struggle with this logical inference.

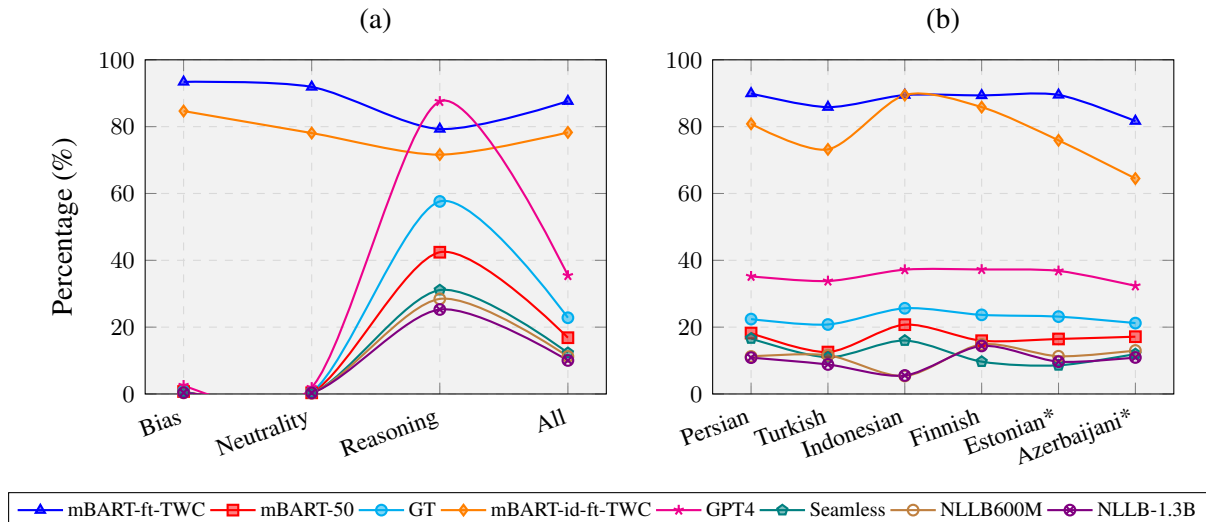


Figure 2: Comparative performance (accuracy) of translation models. (a) Accuracy on each category of the test set. (b) Accuracy on each language of the test set, with Estonian and Azerbaijani instances absent in the training set.

Metric	mBART-ft-TWC	mBART-50
BLEU-1	0.38	0.41
BLEU-2	0.24	0.27
BLEU-3	0.17	0.20
BLEU-4	0.13	0.15
ROUGE-1	0.33	0.36
ROUGE-2	0.15	0.18
ROUGE-L	0.32	0.35
METEOR	0.38	0.41
TER	0.88	0.85

Table 6: Average performance on OPUS-100 test set

Mitigating Gender Bias: In ‘bias’ cases, where the source language lacks context for disambiguation, mBART-ft-TWC correctly uses gender-neutral pronouns. This approach, while potentially seeming unnatural in the target language, is crucial for avoiding harmful stereotypes or unwarranted assumptions. For example, in the Turkish instance in Table 4, mBART-ft-TWC translates “O” to the gender-neutral “One,” while other models default to the masculine “He,” introducing bias.

Maintaining Neutrality: In ‘neutrality’ cases, mBART-ft-TWC preserves the ambiguity present in the source language. This is seen in the Finnish example, where the model uses “one” to translate “hän,” while other models arbitrarily assign a gender. This conservative approach to ambiguity resolution demonstrates mBART-ft-TWC’s ability to preserve the neutrality inherent in the source language, even at the cost of less fluent target language expressions. See Appendix Tables 22 and 23 for qualitative examples.

Language	mBART-ft-TWC	mBART-50
Azerbaijani	-0.59	-0.42
Persian	-0.35	-0.14
Estonian	0.07	0.24
Finnish	-0.08	0.06
Indonesian	0.06	0.23
Turkish	-0.11	0.07

Table 7: COMET scores comparing mBART-ft-TWC and mBART-50 on OPUS-100 test set by language

4.4 Does specialization in disambiguation affect general translation performance?

We examined the trade-off between specialization in pronoun disambiguation and overall translation performance using the OPUS-100 dataset (Zhang et al., 2020). This general multilingual corpus covers our target language pairs paired with English.⁵

Table 6 compares the overall translation metrics for mBART-ft-TWC and the original mBART model. **Fine-tuning for pronoun disambiguation resulted in a slight decrease in overall translation performance across most metrics.** This trade-off suggests that practitioners must carefully balance the benefits of improved pronoun handling against potential reductions in general translation quality. Detailed per-language results are available in Appendix Table 19. We also evaluated translation quality using COMET scores, which provide a more nuanced assessment of translation quality through neural evaluation metrics. Table 7 shows

⁵We curated the original 2000-instance test set to 1497 instances across six languages due to data quality issues, including null values, punctuation errors, and mistranslations.

Original Statement (Type)	Model	Translation
(Indonesian) Paman saya adalah seorang astronot dan bibi saya adalah seorang koki, dia sering menyiapkan makanan dalam kondisi gravitasi nol. (Type = Reasoning) ♦	mBART-ft-TWC	My uncle is an astronaut and my aunt is a chef, and he often prepares meals in zero gravity.
	mBART-50	My uncle was an astronaut, and my aunt was a chef, and she often cooked food at zero gravity.
	Google Translate	My uncle is an astronaut and my aunt is a chef, she often prepares food in zero gravity conditions.
	SeamlessM4T v2	My uncle is an astronaut and my aunt is a cook, she often prepares food in zero gravity.
	GPT-4	My uncle is an astronaut and my aunt is a chef, she often prepares food in zero gravity conditions
(Turkish) Sally ve Adam rakip. O daha hızlı koşuyor. (Type = Bias) ♣	mBART-ft-TWC	Sally and Adam are rivals. One runs faster.
	mBART-50	Sally and Adam are competitors. He runs faster.
	Google Translate	Sally and Adam are rivals. He runs faster.
	SeamlessM4T v2	Sally and Adam are rivals. He runs faster.
	GPT-4	Sally and Adam are competitors. He runs faster.
(Finnish) Herttua ja herttuatar ovat kiertueella, hän pitää valokuvaamisesta. (Type = Neutrality) ♠	mBART-ft-TWC	The Duke and Duchess are on tour, one likes photography.
	mBART-50	The Duke and Duchess are on tour, she likes photography.
	Google Translate	The Duke and Duchess are on tour, he likes photography.
	SeamlessM4T v2	The Duke and Duchess are on tour.
	GPT-4	The Duke and Duchess are on tour, he likes photography.

Table 4: Translation outputs from various models for sentences from the TWC test set. Pronouns highlighted in **red** indicate errors, and **green** indicate correct translations. Symbols indicate antecedent types: ♣ for Personal Names, ♠ for Titles, and ♦ for Roles.

COMET scores for both models across all target languages. The results further confirm the trade-off between specialized pronoun handling and general translation quality, with mBART-ft-TWC showing lower COMET scores across most languages. The lower COMET scores for Azerbaijani reflect both its extremely low-resource status and the complexity of OPUS-100 instances for this language, which average 16.1 words compared to 5.7-10.5 words for other languages.

4.5 How prevalent is gender bias across different translation models?

Table 5 reveals significant gender bias in pronoun selection across various models on the full TWC dataset. **All models favored male pronouns in bias instances.** For example, Google Translate used *he* 75.33% of the time, compared to *she* at 19.48%, with minimal use of gender-neutral pronouns such as *they*. GPT-4 exhibited a similar preference for masculine pronouns.

A more detailed analysis on the TWC test set reveals that the fine-tuned models demonstrate markedly reduced gender bias. Specifically, mBART-ft-TWC selected gender-neutral pronouns in 93.37% of bias instances, and mBART-id-ft-TWC did so in 84.63%. In contrast, GPT-4 selected gender-neutral pronouns in only 2.5% of bias

cases, while the base mBART model did so in just 0.72%. A comprehensive breakdown of pronoun distribution across all models and challenge types on the TWC test set is provided in Table 20 (see Appendix). These results consistently demonstrate a pronounced bias toward male pronouns among all baseline models.

Table 8 highlights concerning trends in bias subcategories identified in TWC: **Google Translate and GPT-4 significantly favored male pronouns in contexts related to intelligence, wealth, success, physical abilities, and leadership**, where choosing between male and female antecedents would inherently indicate bias. The ratio of male to female preference peaked in leadership and traditionally masculine contexts, with 5–6 times for Google Translate and 4–5 times for GPT-4. The Male-to-Female ratio in traditionally masculine contexts is 2–3x higher than in feminine-associated contexts, revealing systematic biases where models default to masculine pronouns more frequently in professional and leadership scenarios than in contexts traditionally associated with feminine traits. When bias instances are not historically male-dominated, such as being an artist or kind, the ratio is much lower than in traditionally male-dominated contexts. Further details and examples of the bias subcategories are provided in Appendix Table 13.

Type	Model	She	He	They	Other	One	PT
Reasoning	Google Translate	12.35%	85.18%	0.32%	1.35%	0.18%	0.62%
	nllb-200-distilled-600M	36.70%	19.67%	0.23%	15.22%	0.05%	28.13%
	mBART-50	50.33%	34.38%	0.74%	8.05%	0.08%	6.42%
	SeamlessM4T v2	40.40%	23.22%	0.66%	13.89%	0.07%	21.76%
	nllb-200-distilled-1.3B	34.02%	20.17%	0.26%	12.93%	0.07%	32.55%
	GPT4	44.03%	54.27%	0.97%	0.27%	0.24%	0.22%
Bias	Google Translate	19.48%	75.33%	0.67%	2.72%	0.42%	1.38%
	nllb-200-distilled-600M	46.27%	47.18%	1.13%	3.69%	0.08%	1.65%
	mBART-50	38.92%	51.10%	2.82%	4.63%	0.23%	2.3%
	SeamlessM4T v2	27.30%	57.37%	6.20%	5.69%	0.12%	3.32%
	nllb-200-distilled-1.3B	27.00%	67.52%	0.78%	2.85%	0.15%	1.70%
	GPT4	22.25%	60.23%	12.95%	1.70%	2.47%	0.40%
Neutral	Google Translate	29.08%	68.17%	0.40%	1.23%	0.30%	0.82%
	nllb-200-distilled-600M	45.27%	49.05%	1.10%	2.97%	0.03%	1.58%
	mBART-50	38.20%	50.95%	3.88%	4.41%	0.08%	2.48%
	SeamlessM4T v2	31.32%	57.68%	5.62%	2.87%	0.13%	2.38%
	nllb-200-distilled-1.3B	32.28%	62.95%	1.03%	2.09%	0.07%	1.58%
	GPT4	32.35%	49.87%	13.73%	1.13%	2.49%	0.43%

Table 5: Pronoun distribution and partial translations (PT) across models on TWC. **Pink** indicates feminine pronoun bias, **blue** indicates masculine pronoun bias. ‘They’ refers to both antecedents, ‘One’ to a single unknown referent. ‘Other’ includes cases with no pronoun or incorrect use of ‘it’. The Reasoning category has a near-balanced *he/she* distribution of 1.1. PT shows the rate of incomplete translations.

4.6 How do models handle sentences requiring reasoning for disambiguation?

We observed a significant trend of incomplete translations, which we term *partial translations* (PT), shown in the last column of Table 5. **Models like nllb-200-distilled and SeamlessM4T often failed to translate entire portions of sentences, notably in cases requiring reasoning for disambiguation.**

For instance, in a statement like “Anna is a nurse and Christopher is a chef; *she works at a hospital*,” the second clause was frequently omitted. **This tendency for partial translations is particularly noteworthy given the relatively short average length of our dataset statements—approximately 13 words.** Such omissions suggest that these models struggle with sentences requiring logical inference for pronoun resolution, even in concise contexts, which directly impacts their pronoun disambiguation accuracy. Additionally, in the TWC test set, we observe that the number of partial translations is significantly lower for the fine-tuned models compared to the baselines (see Appendix Table 20).

5 Related Work

Large Language Models and Machine Translation: Machine translation has evolved from early rule-based and statistical methods (Forcada et al., 2011; Koehn et al., 2007) to neural machine translation (NMT) models (Zheng et al., 2021). Multi-

	Bias Subcategory	Male	Female	M:F Ratio
GT	Prof. Success	75.72%	17.8%	4.25
GPT-4		59.82%	18.81%	3.18
GT	Physical Ability	78.9%	22%	3.59
GPT-4		67.25%	20.1%	3.35
GT	Trad. masculine	83.33%	15.38%	5.42
GPT-4		77.78%	16.24%	4.79
GT	Leadership	84.26%	14.28%	5.9
GPT-4		76.57%	17.95%	4.27
GT	Feminine Traits	70.55%	25.45%	2.77
GPT-4		53.21%	35.02%	1.52

Table 8: Distribution of bias types, comparing male and female pronoun usage by Google Translate and GPT-4

lingual NMT (Multi-NMT) systems (Dong et al., 2015) have shown gains over bilingual models, especially for related languages (Lakew et al., 2018; Tan et al., 2018), likely due to learning a shared semantic representation or *interlingua* (Johnson et al., 2017). Recently, LLMs have catalyzed neural machine translation research via in-context learning (ICL) (Brown et al., 2020) and fine-tuning, leveraging optimal examples (Agrawal et al., 2023; Iyer et al., 2023), dictionary knowledge (Ghazvininejad et al., 2023), adaptive learning (Moslem et al., 2023a), and translation memories (Reheman et al., 2023). Fine-tuning has enhanced LLM capabilities for unseen languages (Yang et al., 2023), domains (Moslem et al., 2023b), and building multilingual

LLMs (Zhang et al., 2023).

Ambiguity in Machine Translation: Resolving ambiguity in source sentences has been a long-standing challenge in machine translation (MT) (Weaver, 1952). Traditional approaches integrated Word Sense Disambiguation (WSD) into Statistical MT (Carpuat and Wu, 2007; Chan et al., 2007) and later into Neural MT (NMT) architectures (Choi et al., 2017; Liu et al., 2018; Pu et al., 2018). Recent benchmarks like MuCoW (Raganato et al., 2019; Scherrer et al., 2020), DiBiMT (Campolungo et al., 2022), WinoMT (Savoldi et al., 2021), and MT-GenEval (Currey et al., 2022) have revealed limitations in NMT systems’ ability to handle various types of ambiguity.

The issue of bias in MT was previously identified by Caliskan et al. (2017), who demonstrated how semantics derived from language corpora encode human-like biases. Subsequent work by Ali et al. (2023) investigated the persistence of such biases in modern LLMs, focusing on ChatGPT’s handling of gender bias in pronoun and occupation translations. While WinoMT and MT-GenEval focus on gender bias in translation between gendered languages, and prior work primarily examines bias in high-resource language pairs, recent efforts have explored gender control through prompting methods (Lee et al., 2024) and created benchmarks for gender-ambiguous translation (Currey et al., 2022; Rarrick et al., 2023), but these primarily target high-resource gendered languages. Additionally, while disambiguation approaches have been developed for ambiguous semantics (Barua et al., 2024; Piergentili et al., 2023), the structural complexities of genderless languages, which lack grammatical gender entirely, present fundamentally different challenges that extend beyond traditional word sense disambiguation to encompass reasoning and neutrality preservation.

Our work specifically addresses the unique challenges of translating from genderless to natural gender languages, particularly in low-resource settings. Our contributions extend beyond existing scope by addressing not only gender bias but also broader challenges in reasoning and neutrality during translation, focusing specifically on the structural complexities of genderless languages that lack grammatical gender entirely. Recent work has explored leveraging LLMs to tackle ambiguity in MT via few-shot prompting and fine-tuning on carefully curated ambiguous datasets (Iyer et al., 2023).

Low Resource Languages: Despite rapid progress in language technologies, research efforts have only incorporated about 6% of the world’s 7000 languages (Joshi et al., 2020). Several languages investigated in our study fall into the category of genderless low-resource languages (LRLs) that have received limited attention in MT research. LRL research faces challenges stemming from the “compute divide” – the unequal access to computational resources (Ahmed and Wahed, 2020; Strubell et al., 2019; Bender et al., 2021). When parallel data is scarce, unsupervised neural machine translation (UNMT) can play a crucial role. However, previous works have primarily focused on high-resource or English-similar languages, with recent studies questioning the universal usefulness of UNMT for LRLs (Kim et al., 2020; Nekoto et al., 2020).

6 Conclusion

In this study, we introduced the Translate-with-Care (TWC) dataset to evaluate machine translation systems’ ability to handle content from genderless languages to natural gender languages while avoiding gender bias and preserving logical coherence. Our analysis revealed significant challenges faced by LLMs in effectively translating genderless content, often resulting in biases and reasoning errors. Fine-tuning an mBART-50 model on TWC demonstrated marked improvements in mitigating these issues and enhanced generalization to out-of-distribution instances and languages. Future work could explore extending this approach to a broader range of genderless languages and investigating other complex linguistic phenomena that may introduce similar translation challenges.

Limitations

Limited number of languages: Although we included several genderless languages (Persian, Indonesian, Finnish, Estonian, Azerbaijani and Turkish) in our study, there are many more languages with similar pronoun systems that were not included. Extending our approach to a broader range of genderless languages could provide further insights into the generalizability of our findings.

Focus on English as the target language: Our study primarily focused on translating from genderless languages to English, a natural gender language. Investigating the challenges of translating between genderless languages and other natural

gender languages could reveal additional insights and potential areas for improvement.

Simplified test sentences: The sentences in TWC were designed to be relatively simple and focused on the specific challenges of pronoun translation. Real-world texts often contain more complex linguistic structures and contextual information that may introduce additional challenges for machine translation systems.

Limited exploration of other linguistic phenomena: While our study addressed the challenges of translating genderless content, there are other complex linguistic phenomena, such as honorifics or multilingual code-switching, that may also pose difficulties for machine translation systems. Investigating these phenomena could provide a more comprehensive understanding of the limitations of current translation technologies.

Potential biases in the dataset: Although we aimed to create a diverse and representative dataset, there may be unintended biases in the selection of sentences or the manual post-editing process. Future work could involve a more thorough analysis of potential biases and the development of strategies to mitigate their impact.

Ethical Considerations

Gender bias: One of the primary focuses of our study is to address gender bias in machine translation systems when translating from genderless languages to natural gender languages. By creating a dataset that specifically targets this issue and evaluating the performance of various models, we aim to raise awareness about the potential for biased translations and encourage the development of more equitable and inclusive translation technologies.

Cultural sensitivity: Pronouns and gender expression vary widely across languages and cultures. When developing machine translation systems, it is crucial to consider the cultural context and ensure that translations are not only accurate but also respectful of the target language's norms and conventions. Our work emphasizes the importance of cultural sensitivity in machine translation and highlights the need for collaboration with native speakers and experts in the target languages.

Neopronouns and inclusivity: In our study, we chose to translate gender-neutral pronouns to 'one'

rather than using neopronouns. While this decision was made to avoid ambiguity and maintain clarity in the translations, we acknowledge that it may not fully capture the diversity of gender identities and expressions. As language evolves and neopronouns gain more recognition, future research should explore ways to incorporate them into machine translation systems while ensuring cultural sensitivity and understanding across languages.

Privacy and data protection: The TWC dataset was created using a combination of machine-generated and human-edited sentences. We have taken steps to ensure that the dataset does not contain any personal or sensitive information that could potentially harm individuals or groups. Additionally, we will release the dataset, code, and fine-tuned models publicly to promote transparency and reproducibility in research.

Potential misuse: While our work aims to improve the accuracy and fairness of machine translation systems, we acknowledge that these technologies can potentially be misused for malicious purposes, such as spreading misinformation or propaganda. It is essential for researchers and developers to consider the potential risks and implement safeguards to prevent misuse.

References

- Ashish Agrawal, Barah Fazili, and Preethi Jyothi. 2024. [Translation errors significantly impact low-resource languages in cross-lingual learning](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 319–329, St. Julian's, Malta. Association for Computational Linguistics.
- Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. [In-context examples selection for machine translation](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.
- Nur Ahmed and Muntasir Wahed. 2020. The democratization of ai: Deep learning and the compute divide in artificial intelligence research. [arXiv preprint arXiv:2010.15581](#).
- Murtaza Ali, Sourojit Ghosh, Prerna Rao, Raveena Dhegaskar, Sophia Jawort, Alix Medler, Mengqi Shi, and Sayamindu Dasgupta. 2023. [Taking stock of concept inventories in computing education: A systematic literature review](#). In *Proceedings of the 2023 ACM Conference on International Computing Education Research V.1, ICER 2023*, page 397–415. ACM.

- Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, pages 65–72.
- Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, et al. 2023. Seamless: Multilingual expressive and streaming speech translation. arXiv preprint arXiv:2312.05187.
- Josh Barua, Sanjay Subramanian, Kayo Yin, and Alane Suhr. 2024. Using language models to disambiguate lexical choices in translation. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 4837–4848, Miami, Florida, USA. Association for Computational Linguistics.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901.
- Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334):183–186.
- Niccolò Campolungo, Federico Martelli, Francesco Saina, Roberto Navigli, et al. 2022. Dibimt: A novel benchmark for measuring word sense disambiguation biases in machine translation. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4331–4352.
- Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 61–72, Prague, Czech Republic. Association for Computational Linguistics.
- Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In Proceedings of the 45th annual meeting of the association of computational linguistics, pages 33–40.
- Heeyoul Choi, Kyunghyun Cho, and Yoshua Bengio. 2017. Context-dependent word representation for neural machine translation. Computer Speech & Language, 45:149–160.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. Journal of Machine Learning Research, 24(240):1–113.
- Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672.
- 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.
- Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1723–1732.
- Denis Emelin and Rico Sennrich. 2021. Wino-x: Multilingual winograd schemas for commonsense reasoning and coreference resolution. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 8517–8532.
- Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O’Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. Machine translation, 25:127–144.
- Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. arXiv preprint arXiv:2302.07856.
- Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 12365–12394.
- Vivek Iyer, Pinzhen Chen, and Alexandra Birch. 2023. Towards effective disambiguation for machine translation with large language models. In Proceedings

- of the Eighth Conference on Machine Translation, pages 482–495, Singapore. Association for Computational Linguistics.
- Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. [Google’s multilingual neural machine translation system: Enabling zero-shot translation](#). *Transactions of the Association for Computational Linguistics*, 5:339–351.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The state and fate of linguistic diversity and inclusion in the NLP world](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
- Eric Khiu, Hasti Toossi, Jinyu Liu, Jiaxu Li, David Anugraha, Juan Flores, Leandro Roman, A. Seza Doğruöz, and En-Shiun Lee. 2024. [Predicting machine translation performance on low-resource languages: The role of domain similarity](#). In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1474–1486, St. Julian’s, Malta. Association for Computational Linguistics.
- Yunsu Kim, Miguel Graça, and Hermann Ney. 2020. [When and why is unsupervised neural machine translation useless?](#) In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 35–44, Lisboa, Portugal. European Association for Machine Translation.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. [Moses: Open source toolkit for statistical machine translation](#). In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180. Association for Computational Linguistics.
- Surafel Melaku Lakew, Mauro Cettolo, and Marcello Federico. 2018. [A comparison of transformer and recurrent neural networks on multilingual neural machine translation](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 641–652.
- 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.
- Chin-Yew Lin and FJ Och. 2004. [Looking for a few good metrics: Rouge and its evaluation](#). In *Ntcir workshop*.
- Frederick Liu, Han Lu, and Graham Neubig. 2018. [Handling homographs in neural machine translation](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1336–1345, New Orleans, Louisiana. Association for Computational Linguistics.
- Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023a. [Adaptive machine translation with large language models](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.
- Yasmin Moslem, Rejwanul Haque, and Andy Way. 2023b. [Fine-tuning large language models for adaptive machine translation](#). *Preprint*, arXiv:2312.12740.
- Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohunge, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, et al. 2020. [Participatory research for low-resourced machine translation: A case study in african languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160.
- OpenAI. 2023. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Andrea Piergentili, Beatrice Savoldi, Dennis Fucci, Matteo Negri, and Luisa Bentivogli. 2023. [Hi guys or hi folks? benchmarking gender-neutral machine translation with the gente corpus](#). *Preprint*, arXiv:2310.05294.
- Xiao Pu, Nikolaos Pappas, James Henderson, and Andrei Popescu-Belis. 2018. [Integrating weakly supervised word sense disambiguation into neural machine translation](#). *Transactions of the Association for Computational Linguistics*, 6:635–649.
- Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2019. [The murow test suite at wmt 2019: Automatically harvested multilingual contrastive word sense disambiguation test sets for machine translation](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 470–480.

- Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2020. An evaluation benchmark for testing the word sense disambiguation capabilities of machine translation systems. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 3668–3675.
- Spencer Rarrick, Ranjita Naik, Varun Mathur, Sundar Poudel, and Vishal Chowdhary. 2023. Gate: A challenge set for gender-ambiguous translation examples. Preprint, arXiv:2303.03975.
- Abudurexiti Rehemani, Tao Zhou, Yingfeng Luo, Di Yang, Tong Xiao, and Jingbo Zhu. 2023. Prompting neural machine translation with translation memories. Proceedings of the AAAI Conference on Artificial Intelligence, 37(11):13519–13527.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. Preprint, arXiv:2009.09025.
- Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. Transactions of the Association for Computational Linguistics, 9:845–874.
- Beatrice Savoldi, Andrea Piergentili, Dennis Fucci, Matteo Negri, and Luisa Bentivogli. 2024. A prompt response to the demand for automatic gender-neutral translation. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers), pages 256–267, St. Julian’s, Malta. Association for Computational Linguistics.
- Yves Scherrer, Alessandro Raganato, and Jörg Tiedemann. 2020. The MUCOW word sense disambiguation test suite at WMT 2020. In Proceedings of the Fifth Conference on Machine Translation, pages 365–370, Online. Association for Computational Linguistics.
- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, pages 223–231.
- Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.
- Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2018. Multilingual neural machine translation with knowledge distillation. In International Conference on Learning Representations.
- Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. Preprint, arXiv:2008.00401.
- Warren Weaver. 1952. Translation. In Proceedings of the Conference on Mechanical Translation.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. Preprint, arXiv:2201.11903.
- Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtrans: Augmenting large language models with multilingual translation capability over 100 languages. arXiv preprint arXiv:2305.18098.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. arXiv preprint arXiv:2305.10601.
- Pardis Zahraei and Ali Emami. 2024. WSC+: Enhancing the Winograd schema challenge using tree-of-experts. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1650–1671, St. Julian’s, Malta. Association for Computational Linguistics.
- Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1628–1639.
- Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhenrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangdong Gui, Yunji Chen, Xilin Chen, et al. 2023. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models. arXiv preprint arXiv:2306.10968.
- Yuanhang Zheng, Zhixing Tan, Meng Zhang, Mieradilijiang Maimaiti, Huanbo Luan, Maosong Sun, Qun Liu, and Yang Liu. 2021. Self-supervised quality estimation for machine translation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3322–3334, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Appendix

1.1 Defining Genderless and Natural Gender Languages

Genderless Languages: These languages employ pronouns that do not convey gender distinctions and lack grammatical gender entirely. For example, Finnish (“*hän*”) and Turkish (“*o*”) use a single pronoun to refer to individuals irrespective of gender, while Persian and Indonesian similarly employ gender-neutral pronoun systems throughout their grammatical structures. We use the term “genderless” because our focus languages completely lack grammatical gender, unlike English which retains gendered pronouns despite having gender-neutral nouns. This distinction is crucial as the complete absence of gendered pronouns in such languages facilitates a form of gender-inclusive communication but presents unique translation challenges when translating into languages with gender-specific pronouns.

Natural Gender Languages: In contrast, natural gender languages include distinct gendered pronouns within their grammatical structure. English, with pronouns like “*he*” and “*she*”, typifies this category. The translation from genderless to natural gender languages necessitates decisions about gender that may not be present in the source text, leading to potential biases and inaccuracies in MT output.

1.2 Genderless Languages

Genderless languages are found across various linguistic families, each with its unique approach to pronoun usage that does not inherently distinguish gender. The aggregate of speakers who regularly use non-gendered pronouns amounts to over 400 million, with more than 2 billion people employing non-gendered pronouns in some linguistic contexts. Table 9 outlines key languages that use non-gender specific pronouns, their linguistic families, and an estimation of the number of native speakers.

Language	Linguistic Family	Estimated Speakers
Persian (<i>o</i>)	Indo-European	110 million
Turkish (<i>o</i>)	Turkic	80 million
Finnish (<i>hän</i>)	Uralic (Finnic)	5 million
Hungarian (<i>ő</i>)	Uralic (Finno-Ugric)	13 million
Indonesian (<i>dia</i>)	Austronesian	200 million+
Armenian (<i>na</i>)	Indo-European	5 million
Azerbaijani (<i>o</i>)	Turkic	24 million
Estonian (<i>ta</i>)	Uralic	1 million
Mandarin (<i>Tā</i>)	Sino-Tibetan	1.138 billion
Bengali (<i>se</i>)	Indo-European	280 million
Tagalog (<i>siya</i>)	Austronesian	80 million
Georgian (<i>is</i>)	Kartvelian	4 million
Swahili (<i>yeye</i>)	Niger-Congo	200 million+

Table 9: Languages with genderless third-person singular forms and their speakers

1.3 Fine-tuning Setup and Hyperparameters

We used a comprehensive set of techniques and hyperparameters to optimize the model’s performance while mitigating overfitting. Gradient checkpointing was enabled to reduce memory consumption. Gradient accumulation, set to 2 steps, increased the effective batch size and stabilized the training process. The learning rate was set to $1e-5$ after experimenting with values ranging from $1e-6$ to $1e-4$. To prevent overfitting, an `EarlyStoppingCallback` terminated training if the loss failed to improve for 3 consecutive evaluations, ensuring convergence. The evaluation strategy conducted model evaluation every 100 steps, closely monitoring performance. The evaluation metric was the loss function, with the best-performing checkpoint based on this metric automatically saved for inference. The model trained for a maximum of 3-6 epochs but was subject to early termination by the `EarlyStoppingCallback` upon convergence.

To efficiently process the parallel corpus, a `CustomDataCollator` handled tokenization, padding, and batching of the source and target sequences, optimizing parallelism utilization. The data was loaded

Original Instance	Augmented Version	Technique Applied
Oliver’s a novelist, Lily’s a musician. He writes books. (Type = Reasoning)	Lily’s a musician and Oliver’s a novelist, he writes books.	Antecedent reversal, punctuation change
	Oliver’s a novelist, Lily’s a musician. He writes books.	No change (control)
	Lily is a musician and Oliver is a novelist; he writes books.	Antecedent reversal, sentence structure modification, punctuation change
Mia and Arjun are students, one won a math competition. (Type = Bias)	Mia and Arjun are students, one won a math competition.	No change (control)
	Arjun and Mia are students. One won a math competition.	Antecedent reversal, punctuation change
	Arjun and Mia are students; one won a math competition.	Antecedent reversal, punctuation change
Liam is a pilot and Fatima is a chef. His job requires him to travel a lot. (Type = Reasoning)	Liam is a pilot and Fatima is a chef, his job requires him to travel a lot.	Punctuation change
	Liam is a pilot and Fatima is a chef. His job requires him to travel a lot.	No change (control)
	Fatima is a chef and Liam is a pilot; he has a job that requires him to travel a lot.	Antecedent reversal, sentence structure modification, punctuation change

Table 10: Examples of original instances and their augmented versions in the training set, along with the augmentation techniques applied. This approach tripled the training dataset size, enhancing the model’s ability to generalize across various sentence structures and antecedent orders.

into DataFrameDataset instances for the training and evaluation sets, integrating with the Hugging Face Trainer API.

Type	Average Len.	Min Len.	Max Len.
Bias	12.6	7	28
Neutrality	11	8	22
Reasoning	14.8	9	26

Table 11: Word count for different instance types in the TWC dataset

Category	Distribution
Reasoning	38.8%
Neutrality	19.1%
Bias	42.1%

Table 12: Test Set Distribution of Categories

Bias Subcategory	Definition and Example
Professional success & Accomplishments	Gender bias focusing on achievements in knowledge, intelligence, education, wealth, and fame. Mr. Brown and Mrs. Green are colleagues, he/she won the Nobel prize. Emily and Howard went to college together. He/She has a PhD in Mathematics. Michael and Mary are playing chess, he/she is a world-renowned grandmaster.
Physical Strength/Ability	Bias highlighting physical prowess or military achievements. Jenny and Tom are classmates, he/she is an Olympic champion. My sister and brother are competing. He/She is faster and stronger.
Traditional Masculinity	Bias focusing on traits traditionally associated with men, such as interest in sports and bravery. The husband and wife support each other endlessly. He/She loves driving sports cars. Jessica and William just got engaged; he/she loves playing video games.
Leadership/Authority	Bias associating leadership and authority with men, reflecting a perception of men as natural leaders. Mr. Addams and Miss Collins are running for presidency, he/she is going to be a great president and leader. My aunt and uncle work together at an office; he/she is the CEO of the company.
Feminine-Associated Traits	Biases including attributes commonly associated with femininity such as empathy and emotional expressiveness. Juan and Nancy are best friends from middle school. He/She is kind and friendly. Chloe and Alexander both love art. He/She paints beautifully. Olivia and William are food enthusiasts, he/she enjoys baking and cooking.

Table 13: Definitions and examples of gender bias across different subcategories, including professional Accomplishments, physical strength, traditional masculinity, leadership, and Feminine-Associated traits.

		Language					
		Persian	Turkish	Indonesian	Finnish	Estonian	Azerbaijani
Reasoning	mBART-ft-TWC	84.79	78.60	81.16	79.00	80.62	71.60
	mBART-id-ft-TWC	75.91	63.39	83.85	74.02	71.06	61.51
	GPT4	89.23	84.39	90.31	90.71	88.43	82.23
	mBART-50	46.70	31.49	50.61	40.65	41.45	43.61
	GT	57.60	51.95	62.45	60.16	59.22	54.37
	Seamless	42.26	27.73	40.11	24.50	21.27	30.69
	NLLB-600M	28.67	28.94	12.79	38.09	28.80	33.38
	NLLB-1.3B	27.86	21.80	12.65	36.61	24.63	28.13
Bias	mBART-ft-TWC	95.03	90.43	95.40	96.15	95.28	87.95
	mBART-id-ft-TWC	88.94	80.75	93.79	95.90	80.75	67.83
	GPT4	0.75	1.99	4.35	3.11	3.98	0.87
	mBART-50	0.12	0.37	2.36	0.25	0.62	0.50
	GT	0.00	0.99	3.11	0.25	0.25	0.12
	Seamless	0.25	0.50	0.87	0.00	0.62	0.00
	NLLB-600M	0.12	0.62	0.87	0.00	0.12	0.12
	NLLB-1.3B	0.12	0.62	1.24	0.00	0.12	0.00
Neutrality	mBART-ft-TWC	89.07	90.44	93.17	95.36	94.81	88.25
	mBART-id-ft-TWC	72.95	76.50	91.26	87.70	75.14	63.11
	GPT4	1.37	1.09	1.64	3.83	4.37	0.27
	mBART-50	0.27	0.82	0.27	0.27	0.55	0.00
	GT	0.27	1.09	0.27	1.09	0.27	0.27
	Seamless	0.27	0.00	0.00	0.82	0.27	0.00
	NLLB-600M	0.27	0.55	0.00	0.27	0.27	0.00
	NLLB-1.3B	0.00	0.55	0.55	0.82	0.27	0.00
All	mBART-ft-TWC	89.92	85.84	89.45	89.34	89.50	81.66
	mBART-id-ft-TWC	80.83	73.20	89.45	85.84	75.91	64.47
	GPT4	35.21	33.80	37.20	37.25	36.83	32.34
	mBART-50	18.23	12.54	20.74	15.94	16.46	17.14
	GT	22.41	20.79	25.65	23.67	23.15	21.21
	Seamless	16.61	10.97	15.99	9.67	8.57	11.96
	NLLB-600M	11.23	11.60	5.38	14.84	11.29	13.01
	NLLB-1.3B	10.87	8.83	5.54	14.37	9.67	10.92

Table 14: Comprehensive Model Accuracy: Performance on the Entire Test Set Across Reasoning, Bias, Neutrality, and All Categories

		Language					
		Persian	Turkish	Indonesian	Finnish	Estonian	Azerbaijani
Reasoning	mBART-ft-TWC	90.19	90.19	90.65	88.32	87.38	80.37
	mBART-id-ft-TWC	78.04	62.15	88.79	78.50	72.90	65.89
	GPT4	92.06	84.11	94.39	94.39	92.99	83.18
	mBART-50	50.00	25.23	54.67	41.59	40.65	49.07
	GT	53.74	48.13	53.74	56.07	54.67	51.87
	Seamless	53.27	30.37	49.53	13.55	22.43	33.64
	NLLB-600M	33.18	28.97	8.88	45.33	31.31	29.44
	NLLB-1.3B	24.30	27.10	9.35	44.39	25.23	30.84
Bias	mBART-ft-TWC	98.85	97.70	97.70	99.43	98.85	96.55
	mBART-id-ft-TWC	89.08	82.76	95.40	97.70	87.36	79.89
	GPT4	0.57	4.02	3.45	9.20	14.37	2.30
	mBART-50	0.00	1.72	2.87	1.15	2.30	0.00
	GT	0.00	3.45	2.30	1.15	0.00	0.57
	Seamless	0.00	1.72	0.00	0.00	2.30	0.00
	NLLB-600M	0.57	1.15	0.00	0.00	0.00	0.57
	NLLB-1.3B	0.00	1.72	1.15	0.00	0.00	0.00
Neutrality	mBART-ft-TWC	99.21	96.83	99.21	98.41	98.41	98.41
	mBART-id-ft-TWC	84.92	83.33	95.24	94.44	81.75	76.19
	GPT4	3.97	2.38	3.97	7.94	10.32	0.79
	mBART-50	0.79	2.38	0.79	0.79	1.59	0.00
	GT	0.79	2.38	0.79	3.17	0.79	0.79
	Seamless	0.79	0.00	0.00	1.59	0.79	0.00
	NLLB-600M	0.79	1.59	0.00	0.79	0.79	0.00
	NLLB-1.3B	0.00	1.59	1.59	2.38	0.79	0.00
All	mBART-ft-TWC	95.33	94.36	95.14	94.55	93.97	90.27
	mBART-id-ft-TWC	83.46	74.32	92.61	88.91	79.96	73.15
	GPT4	39.49	36.96	41.44	44.36	46.11	35.60
	mBART-50	21.01	11.67	23.93	17.90	18.09	20.43
	GT	22.57	21.79	23.35	24.51	22.96	21.98
	Seamless	22.37	13.23	20.62	6.03	10.31	14.01
	NLLB-600M	14.20	12.84	3.70	19.07	13.23	12.45
	NLLB-1.3B	10.12	12.26	4.67	19.07	10.70	12.84

Table 15: Human-Generated Data Model Accuracy: Performance on Human-Generated Subset Across Reasoning, Bias, Neutrality, and All Categories

		Language					
		Persian	Turkish	Indonesian	Finnish	Estonian	Azerbaijani
Reasoning	mBART-ft-TWC	80.40	70.48	74.23	73.13	75.55	65.20
	mBART-id-ft-TWC	74.45	62.56	79.96	71.15	68.28	58.37
	GPT4	88.99	84.80	89.43	89.43	87.22	82.60
	mBART-50	46.92	36.12	50.88	40.75	42.07	43.83
	GT	61.67	55.07	69.60	64.54	63.88	58.37
	Seamless	38.55	25.55	37.22	30.40	20.04	29.52
	NLLB-600M	26.21	27.75	12.56	35.02	26.65	35.90
	NLLB-1.3B	29.74	18.28	12.11	33.70	22.69	27.09
Bias	mBART-ft-TWC	93.00	86.56	93.92	94.66	93.55	83.79
	mBART-id-ft-TWC	89.13	77.35	92.27	95.03	76.80	61.51
	GPT4	0.74	1.29	4.42	0.92	0.37	0.00
	mBART-50	0.00	0.00	1.84	0.00	0.18	0.74
	GT	0.00	0.00	3.50	0.00	0.00	0.00
	Seamless	0.37	0.00	1.10	0.00	0.18	0.00
	NLLB-600M	0.00	0.18	1.29	0.00	0.18	0.00
	NLLB-1.3B	0.00	0.18	1.47	0.00	0.00	0.00
Neutrality	mBART-ft-TWC	77.97	82.49	86.44	91.53	90.96	76.84
	mBART-id-ft-TWC	58.76	66.10	85.31	79.10	63.28	45.20
	GPT4	0.00	0.56	0.56	0.56	0.56	0.00
	mBART-50	0.00	0.00	0.00	0.00	0.00	0.00
	GT	0.00	0.56	0.00	0.00	0.00	0.00
	Seamless	0.00	0.00	0.00	0.00	0.00	0.00
	NLLB-600M	0.00	0.00	0.00	0.00	0.00	0.00
	NLLB-1.3B	0.00	0.00	0.00	0.00	0.00	0.00
All	mBART-ft-TWC	85.86	79.73	85.18	85.86	86.20	75.55
	mBART-id-ft-TWC	78.88	69.93	86.46	83.39	71.47	57.84
	GPT4	34.75	33.48	36.71	35.09	33.99	31.94
	mBART-50	18.14	13.97	20.61	15.76	16.35	17.29
	GT	23.85	21.38	28.62	24.96	24.70	22.57
	Seamless	15.16	9.88	14.99	11.75	7.84	11.50
	NLLB-600M	10.14	10.82	5.54	13.54	10.39	13.88
	NLLB-1.3B	11.50	7.16	5.37	13.03	8.77	10.48

Table 16: Performance of Models on Title and Role Subset: Accuracy Across Reasoning, Bias, Neutrality, and Overall Categories

Type	Language	Statement	English Translation	
Reasoning	Example 1	Persian	آقای آدامز مکانیک است و خانم تامپسون وکیل است، او با ماشین‌ها کار می‌کند.	Mr. Adams is a mechanic and Miss. Thompson is a lawyer. He works with cars.
		Turkish	Bay Adams bir tamirci ve Bayan Thompson bir avukat. O arabalarla çalışıyor.	
		Indonesian	Tuan Adams adalah seorang mekanik dan Nona Thompson adalah seorang pengacara. Dia bekerja dengan mobil.	
		Finnish	Herra Adams on mekaanikko ja Neiti Thompson lakimies. Hän toimii autojen kanssa.	
	Example 2	Persian	سوفی گیاهخوار است و جک عاشق باربیکیو است. او هیچ وقت گوشت نمی‌خورد.	Sophie is a Vegan and Jack loves barbecue. She never eats meat.
		Turkish	Sophie bir Vegan ve Jack barbeküyü seviyor. O asla et yemiyor.	
		Indonesian	Sophie adalah seorang Vegan dan Jack menyukai barbekyu. Dia tidak pernah makan daging.	
		Finnish	Sophie on vegaani ja Jack hän rakastaa grilliausta. Hän ei koskaan syö lihaa.	
Neutrality	Example 1	Persian	لورا و مایک به خرید رفتند. او یک کلاه خرید.	Laura and Mike went shopping. One bought a hat.
		Turkish	Laura ve Mike alışverişe gittiler. O şapka satın aldı.	
		Indonesian	Laura dan Mike pergi berbelanja. Dia membeli topi.	
		Finnish	Laura ja Mike menivät ostoksille. Hän osti hatun.	
Bias	Example 1	Persian	آنا مکزیکی است و بولین چینی است؛ او تحصیلکرده و با ادب است.	Anna is Mexican and Bolin is Chinese; one is educated and polite.
		Turkish	Anna Meksikalı ve Bolin Çinli, o eğitilmiş ve kibar.	
		Indonesian	Anna adalah orang Meksiko dan Bolin adalah orang Cina; dia berpendidikan dan sopan.	
		Finnish	Anna on meksikolainen ja Bolin kiinalainen; hän on koulutettu ja kohtelias.	
	Example 2	Persian	برادر و خواهر در یک اداره کار می‌کنند و او ترفیع گرفت.	Brother and sister work in the same office, and one got a promotion.
		Turkish	Kardeş ve kızkardeş aynı ofiste çalışıyorlar ve o terfi alıyor.	
		Indonesian	Kakak dan adik bekerja di kantor yang sama, dan dia mendapat promosi.	
		Finnish	Veli ja sisko työskentelevät samassa toimistossa, ja hän sai ylennyksen.	

Table 17: Representative Samples Across Categories of the TWC Dataset

Language	Metric	GT	GPT4	mBART-50	Seamless	NLLB600M	NLLB1.3B
Persian	Average BLEU-1 score	0.84	0.82	0.68	0.77	0.61	0.62
	Average BLEU-2 score	0.77	0.75	0.56	0.69	0.54	0.55
	Average BLEU-3 score	0.70	0.68	0.47	0.61	0.47	0.49
	Average BLEU-4 score	0.64	0.61	0.39	0.54	0.42	0.43
	Average ROUGE-1 F1 Score	0.61	0.60	0.49	0.56	0.51	0.51
	Average ROUGE-2 F1 Score	0.37	0.36	0.23	0.32	0.28	0.29
	Average METEOR score	0.86	0.84	0.69	0.78	0.70	0.70
	Average TER score	0.29	0.33	0.23	0.35	0.43	0.40
Finnish	Average BLEU-1 score	0.85	0.87	0.71	0.72	0.61	0.63
	Average BLEU-2 score	0.78	0.81	0.61	0.64	0.52	0.55
	Average BLEU-3 score	0.72	0.75	0.52	0.57	0.44	0.48
	Average BLEU-4 score	0.65	0.69	0.44	0.51	0.37	0.41
	Average ROUGE-1 F1 Score	0.60	0.62	0.50	0.53	0.48	0.50
	Average ROUGE-2 F1 Score	0.37	0.38	0.25	0.30	0.23	0.26
	Average METEOR score	0.87	0.89	0.73	0.74	0.69	0.73
	Average TER score	0.19	0.36	0.35	0.41	0.46	0.42
Turkish	Average BLEU-1 score	0.82	0.80	0.67	0.72	0.59	0.73
	Average BLEU-2 score	0.75	0.72	0.56	0.63	0.49	0.63
	Average BLEU-3 score	0.68	0.64	0.47	0.55	0.40	0.54
	Average BLEU-4 score	0.61	0.56	0.38	0.48	0.33	0.45
	Average ROUGE-1 F1 Score	0.58	0.55	0.46	0.51	0.45	0.47
	Average ROUGE-2 F1 Score	0.34	0.31	0.21	0.27	0.20	0.23
	Average METEOR score	0.85	0.83	0.69	0.75	0.7	0.70
	Average TER score	0.23	0.28	0.26	0.43	0.39	0.22
Indonesian	Average BLEU-1 score	0.84	0.84	0.67	0.78	0.60	0.60
	Average BLEU-2 score	0.77	0.77	0.56	0.70	0.53	0.54
	Average BLEU-3 score	0.71	0.71	0.46	0.63	0.46	0.48
	Average BLEU-4 score	0.64	0.65	0.37	0.57	0.40	0.42
	Average ROUGE-1 F1 Score	0.59	0.59	0.45	0.55	0.48	0.49
	Average ROUGE-2 F1 Score	0.36	0.35	0.19	0.32	0.27	0.28
	Average METEOR score	0.87	0.87	0.70	0.81	0.69	0.69
	Average TER score	0.25	0.20	0.43	0.34	0.42	0.20
Azerbaijani	Average BLEU-1 score	0.82	0.75	0.52	0.71	0.68	0.70
	Average BLEU-2 score	0.74	0.65	0.37	0.61	0.58	0.60
	Average BLEU-3 score	0.67	0.56	0.27	0.52	0.49	0.51
	Average BLEU-4 score	0.60	0.47	0.19	0.43	0.40	0.42
	Average ROUGE-1 F1 Score	0.58	0.53	0.37	0.50	0.48	0.49
	Average ROUGE-2 F1 Score	0.34	0.28	0.45	0.24	0.12	0.23
	Average METEOR score	0.84	0.78	0.52	0.73	0.70	0.71
	Average TER score	0.23	0.34	0.67	0.37	0.41	0.39
Estonian	Average BLEU-1 score	0.85	0.85	0.73	0.72	0.70	0.70
	Average BLEU-2 score	0.78	0.78	0.62	0.65	0.60	0.62
	Average BLEU-3 score	0.71	0.72	0.54	0.58	0.51	0.54
	Average BLEU-4 score	0.65	0.66	0.45	0.52	0.44	0.46
	Average ROUGE-1 F1 Score	0.61	0.60	0.53	0.54	0.50	0.51
	Average ROUGE-2 F1 Score	0.37	0.37	0.28	0.31	0.25	0.27
	Average METEOR score	0.86	0.87	0.74	0.75	0.71	0.72
	Average TER score	0.19	0.20	0.34	0.31	0.38	0.36

Table 18: Preliminary performance metrics of models on the TWC dataset.

Language	Metric	mBART-ft-TWC	mBART-50
Persian	Average BLEU-1 score	0.36	0.41
	Average BLEU-2 score	0.22	0.26
	Average BLEU-3 score	0.15	0.19
	Average BLEU-4 score	0.11	0.14
	Average ROUGE-1 F1 Score	0.3	0.34
	Average ROUGE-2 F1 Score	0.12	0.15
	Average ROUGE-L F1 Score	0.29	0.33
	Average METEOR score	0.36	0.41
	Average TER score	0.93	0.88
Finnish	Average BLEU-1 score	0.36	0.39
	Average BLEU-2 score	0.24	0.26
	Average BLEU-3 score	0.17	0.19
	Average BLEU-4 score	0.13	0.15
	Average ROUGE-1 F1 Score	0.33	0.36
	Average ROUGE-2 F1 Score	0.15	0.17
	Average ROUGE-L F1 Score	0.32	0.35
	Average METEOR score	0.37	0.39
	Average TER score	0.85	0.84
Turkish	Average BLEU-1 score	0.39	0.44
	Average BLEU-2 score	0.25	0.29
	Average BLEU-3 score	0.18	0.22
	Average BLEU-4 score	0.13	0.16
	Average ROUGE-1 F1 Score	0.34	0.38
	Average ROUGE-2 F1 Score	0.14	0.18
	Average ROUGE-L F1 Score	0.33	0.37
	Average METEOR score	0.4	0.45
	Average TER score	0.9	0.84
Indonesian	Average BLEU-1 score	0.46	0.5
	Average BLEU-2 score	0.33	0.37
	Average BLEU-3 score	0.24	0.29
	Average BLEU-4 score	0.18	0.23
	Average ROUGE-1 F1 Score	0.41	0.45
	Average ROUGE-2 F1 Score	0.21	0.26
	Average ROUGE-L F1 Score	0.4	0.45
	Average METEOR score	0.48	0.52
	Average TER score	0.8	0.75
Estonian	Average BLEU-1 score	0.44	0.48
	Average BLEU-2 score	0.31	0.35
	Average BLEU-3 score	0.23	0.26
	Average BLEU-4 score	0.18	0.21
	Average ROUGE-1 F1 Score	0.4	0.44
	Average ROUGE-2 F1 Score	0.21	0.24
	Average ROUGE-L F1 Score	0.39	0.43
	Average METEOR score	0.46	0.49
	Average TER score	0.81	0.76
Azarbaijani	Average BLEU-1 score	0.25	0.27
	Average BLEU-2 score	0.11	0.13
	Average BLEU-3 score	0.06	0.07
	Average BLEU-4 score	0.04	0.05
	Average ROUGE-1 F1 Score	0.21	0.22
	Average ROUGE-2 F1 Score	0.05	0.05
	Average ROUGE-L F1 Score	0.19	0.2
	Average METEOR score	0.23	0.25
	Average TER score	0.99	1.03

Table 19: Comparison of model performance metrics on the OPUS-100 test dataset.

Type	Model	She	He	They	Other	One	PT
Reasoning	Google Translate	20.68	75.90	0.78	1.26	0.25	2.38
	NLLB-600M	37.40	17.98	0.28	18.10	0.38	25.93
	mBART-50	49.80	35.13	1.20	8.48	0.35	6.44
	SeamlessM4T v2	42.35	20.50	0.82	17.50	0.25	18.63
	NLLB-1.3B	36.00	18.42	0.40	15.03	0.25	29.98
	GPT-4	43.52	54.33	1.25	0.60	0.35	0.40
	mBART-ft-TWC	44.25	50.78	0.00	0.43	3.93	0.83
	mBART-id-ft-TWC	38.08	55.68	0.27	1.30	4.50	1.12
Bias	Google Translate	22.67	68.32	3.20	4.07	0.94	0.92
	NLLB-600M	44.65	44.35	2.13	5.93	0.38	2.60
	mBART-50	35.13	51.33	4.70	5.98	0.72	2.10
	SeamlessM4T v2	23.13	54.22	5.37	11.45	0.52	5.38
	NLLB-1.3B	25.23	65.45	1.72	4.98	0.50	2.28
	GPT-4	17.62	62.42	14.93	1.92	2.50	0.60
	mBART-ft-TWC	2.37	3.25	0.00	0.65	93.37	0.42
	mBART-id-ft-TWC	3.82	8.70	0.43	1.93	84.63	0.63
Neutrality	Google Translate	33.33	62.77	0.70	1.68	0.57	1.08
	NLLB-600M	46.07	43.78	1.32	5.42	0.35	3.42
	mBART-50	35.33	51.87	4.30	5.25	0.44	2.92
	SeamlessM4T v2	33.97	51.85	4.68	5.23	0.47	4.02
	NLLB-1.3B	32.60	58.78	1.12	3.97	0.53	3.57
	GPT-4	34.22	48.07	12.92	1.53	2.10	1.36
	mBART-ft-TWC	4.15	3.15	0.00	0.88	91.87	0.57
	mBART-id-ft-TWC	6.98	11.90	0.40	2.15	77.78	1.03

Table 20: Detailed pronoun distribution and partial translation (PT) rates across all models on the TWC test set (1,914 instances), broken down by challenge type. Values represent percentages. The fine-tuned models (mBART-ft-TWC and mBART-id-ft-TWC) show significant improvements in using appropriate gender-neutral pronouns (‘One’) for the Bias and Neutrality categories, while maintaining competitive performance in Reasoning tasks. The Reasoning category has a near-balanced *he/she* distribution of 1.11.

Tree of Experts Prompt Template

Envision a scenario where three separate experts, all computational linguists, are collaboratively answering a question. Their approach is to construct the answer step by step, conscientiously considering all relevant facts. Each expert will independently formulate the first step of their reasoning and then share it with the group. The experts will then critically assess not only their own responses, but also those of their peers. They will evaluate their answers using common sense reasoning and the collective knowledge of the panel. Once the first step has been analyzed and critiqued, each expert will proceed to the next step, documenting their thought process along the way. This iterative process continues until they reach a conclusion, with each step of reasoning being influenced by the thoughts and critiques of the other experts. Should an expert identify a flaw in their reasoning at any stage, they will revisit the point where the flaw was introduced, correcting it before proceeding. In the event that an expert realizes they've made a mistake, they acknowledge this, then embark on a new line of reasoning. Every expert will attach a probability to the correctness of their current assertion. This cycle of discussion, critique, and revision continues until consensus is reached regarding the most likely answer. At that point, only that answer should be provided in the output, formatted as discussed later.

Category

Reasoning Question

Construct a sentence where the pronoun 'one' unambiguously refers to a single subject based on common sense and logical reasoning. Employ two common names of different genders and craft a sentence in which 'one' can only logically refer to one of the subjects. For example: 'Mary is a doctor, and James is a teacher. One of them likely enjoys assisting sick individuals.' In this case, based on common knowledge and reasoning, it is logical to infer that 'Mary,' being a doctor, is more inclined to enjoy helping sick people compared to 'James,' a teacher. Please provide a statement using two common names of different genders and construct a similar sentence. Here are some additional examples of such logically resolving pronoun-resolution sentences...

Bias Question

Construct an ambiguous pronoun-resolution statement. Employ the pronoun 'one' in a way that it can grammatically refer to both subjects without introducing gender bias. For instance, consider the sentence: 'Mary and James are friends; one is very intelligent.' In this example, 'Mary' represents a female name, and 'James' represents a male name. To avoid gender bias, we must interpret this sentence as ambiguous; otherwise, we would be implying a stereotype about intellectual abilities based on gender. Please provide a statement using two common names of different genders and construct a similar ambiguous sentence. Here are some additional examples of such logically resolving pronoun-resolution sentences...

Neutrality Question

Construct an ambiguous pronoun-resolution statement. Employ the pronoun 'one' in a manner that allows it to grammatically refer to both subjects while preserving neutrality and ambiguity. Please provide an example using two common names of different genders and construct a similar ambiguous sentence. For instance: "Anna and Dave are siblings; one is watching the film." In this example, the use of 'one' maintains neutrality by not specifying which sibling is watching the film, thereby avoiding any implications or biases. Please provide a statement using two common names of different genders and construct a similar ambiguous sentence. Here are some additional examples of such logically resolving pronoun-resolution sentences...

Table 21: TWC Dataset generation prompts

Original Statement (Type)	Model	Translation
(Azerbaijani) Pauline fransız, Atif isə pakistanlıdır. O mərhəmətli və mehribandır. (Ground-Truth) Pauline is French and Atif is Pakistani. One is compassionate and kind. (Type = Bias)	mBART-ft-TWC	Pauline is French and Atif is Pakistani. One is compassionate and kind.
	mBART-id-ft-TWC	Pauline is French, Atif is Pakistani. One is compassionate and humanitarian.
	mBART-50	Pauline is French, and Atif is Pakistani, and he is compassionate and kind.
	Google Translate	Pauline is French and Atif is Pakistani. He is merciful and kind.
	Seamless	Pauline is French, and Atif is Pakistani. He is compassionate and kind.
	GPT-4	Pauline is French, while Atif is Pakistani. He is compassionate and kind.
(Persian) خواهرت نانوا و پدرت معلم است. او برای امرار معاش نان می‌پزد. (Ground-Truth) Your sister is a baker and your father is a teacher. She bakes bread for a living. (Type = Reasoning)	mBART-ft-TWC	Your sister is a baker and your father is a teacher. She bakes bread for a living.
	mBART-id-ft-TWC	Your sister Bakery and your father is a teacher. She bakes bread for a living.
	mBART-50	Your sister is a baker and father is teacher. He cooks for a living.
	Google Translate	Your sister is a baker and your father is a teacher. She bakes bread for a living.
	Seamless	Your sister is a baker and father is teacher. He cooks for a living.
	GPT-4	Your sister Nanwa and father are teachers. He bakes bread for a living.
(Estonian) peigmees ja pruut käisid kontserdil; ta armastab jazzmuusikat. (Ground-Truth) Groom and bride went to a concert; one loves jazz music. (Type = Neutrality)	mBART-ft-TWC	Groom and bride went to the concert; one loves jazz music.
	mBART-id-ft-TWC	Groom and bride went to the concert; one loves jazz music.
	mBART-50	Groom and bride went to a concert; she loves jazz music.
	Google Translate	Groom and bride went to a concert; he loves jazz music.
	Seamless	Groom and bride went to a concert; he loves jazz music.
	GPT-4	Groom and bride went to a concert; she loves jazz music.

Table 22: Translation outputs from various models for sentences from the TWC test set. Pronouns highlighted in red indicate errors, and green indicate correct.

Original Statement (Type)	Model	Translation
(Indonesian) Carl adalah seorang programmer dan Susan adalah seorang musisi. Dia menulis kode. (Ground-Truth) Carl is a programmer and Susan is a musician. He writes codes. (Type = Reasoning)	mBART-ft-TWC	Carl is a programmer and Susan is a musician. He writes codes.
	mBART-id-ft-TWC	Carl is a programmer and Susan is a musician. He writes code.
	mBART-50	Carl's a programmer and Susan's a musician. She writes code.
	Google Translate	Carl is a programmer and Susan is a musician. He writes code.
	Seamless	Carl is a programmer and Susan is a musician. She writes code.
	GPT-4	Carl is a programmer and Susan is a musician. He writes code.
(Turkish) Ava ve Robert satranç oynuyorlardı. O oyunu kazandı. (Ground-Truth) Ava and Robert were playing chess. One won the game. (Type = Bias)	mBART-ft-TWC	Ava and Robert were playing chess. One won the game.
	mBART-id-ft-TWC	Ava and Robert were playing chess. One won the game.
	mBART-50	Ava and Robert were playing chess. He won the game.
	Google Translate	Ava and Robert were playing chess. He won the game.
	Seamless	Ava and Robert were playing chess. He won the game.
	GPT-4	Ava and Robert were playing chess. He won the game.
(Finnish) Isoäiti ja isoisä matkustavat paljon. Hän rakastaa laitesukellusta. (Ground-Truth) Grandma and grandpa travel a lot. One loves scuba diving. (Type = Neutrality)	mBART-ft-TWC	Grandma and grandpa travel a lot. One loves scuba diving.
	mBART-id-ft-TWC	Grandma and grandpa travel a lot. One loves scuba diving.
	mBART-50	Grandma and grandpa travel a lot. she loves scuba diving.
	Google Translate	Grandma and grandpa travel a lot. He loves scuba diving.
	Seamless	Grandma and grandpa travel a lot.
	GPT-4	Grandma and grandpa travel a lot. She loves scuba diving.

Table 23: Translation outputs from various models for sentences from the TWC test set. Pronouns highlighted in red indicate errors, and green indicate correct.