GENDER1PERSON: Test Suite for estimating gender bias of first-person singular forms

Maja Popović[1], **Ekaterina Lapshinova-Koltunski**[2]
[1] IU University, Berlin, Germany
maja.popovic@iu.org
[2] University of Hildesheim, Germany
lapshinovakoltun@uni-hildesheim.de

Abstract

The GENDER1PERSON test suite is designed to measure gender bias in translating singular first-person forms from English into two Slavic languages, Russian and Serbian. The test suite consists of 1 000 Amazon product reviews, uniformly distributed over 10 different product categories. Bias is measured through a gender score ranging from -100 (all reviews are feminine) to 100 (all reviews are masculine).

The test suite shows that the majority of the systems participating in the WMT-2025 task for these two target languages prefer the masculine writer's gender. There is no single system which is biased towards the feminine variant. Furthermore, for each language pair, there are seven systems that are considered balanced, having the gender scores between -10 and 10.

Finally, the analysis of different products showed that the choice of the writer's gender depends to a large extent on the product. Moreover, it is demonstrated that even the systems with overall balanced scores are actually biased, but in different ways for different product categories.

1 Introduction

While English does not have many morphological forms related to gender, the two target languages do so. In those languages, gender marking exists not only for pronouns and animate nouns, but also for nouns, adjectives, verbs, determiners and numbers. If the text is written in the first-person singular form and no information about the gender of the author is provided, the translator can choose any of the two binary[1] genders. One of the most frequent affected POS tags are adjectives and past or passive participles. For example, "*I am happy that I bought this*" can be translated into Serbian in two ways, "*Srećan/srećna sam što sam ovo kupio/kupila*", depending on the writer's natural gender. This may result in translation errors, mismatches and inconsistencies, as well as in gender bias.

Our test suite is designed to measure bias of this type of gender in translations from English into Russian and Serbian. It consists of a carefully selected set of user reviews about Amazon products, because these texts are written in the first-person form and therefore very convenient. The test suite also enables the analysis of writer's gender depending on the product category. Although currently covering two target languages, it can easily be extended to more languages with similar rules for first-person singular gender.

Our main motivation was the results of our experiments reported in (Popovic and Lapshinova-Koltunski, 2024). We found some interesting tendencies regarding the writer's gender in user reviews of Amazon products from the DiHuTra corpus (Lapshinova-Koltunski et al., 2022). However, this corpus is designed for investigating differences between human and machine translations, but it is not tailored for exploring gender. The corpus is relatively small, only 196 reviews in total, and one third of them do not contain any indicator of the writer's gender. Therefore, the reported results (especially for different products), while interesting, were not fully reliable. The test suite presented in this paper, is much bigger and contains 1000 reviews.

2 Related work

While there is a large portion of work dealing with different types of gender bias, there are not many studies focussing on the first-person constructions. For example, Habash et al. (2019) propose automatic generation of both gender variants for the first person in Arabic NMT translations.

Our test suite also enables analysing bias depen-

[1] In the analysed target languages there are still no non-binary forms.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 353–375
November 8–9, 2025.

dence on product category. Similarly, bias variation was addressed in (Zhao et al., 2017) who reported that data sets for specific tasks (e.g. cooking) contain significant gender bias and, furthermore, models trained on these data sets further amplify existing bias.

Our test suite uses a gender score as a metric. Cho et al. (2019) also proposes a measure of gender bias, however, in a completely different context: the metric measures the relation between gender and positive/negative expressions or occupations.

There exist other test suites. For instance, Stanovsky et al. (2019) design a test suite for evaluating gender bias in MT related to occupational nouns. Their method was developed for eight target languages, including Russian. Vanmassenhove and Monti (2021) create an English–Italian test suite with a focus on the resolution of natural gender. The authors provide word-level gender tags on the English source side and multiple alternative gender translations, where needed, on the Italian target side. Savoldi et al. (2023, 2024) present a test suite to investigate systems' ability to correctly translate the gender of the speaker int he context of occupations and professions (e.g. "*I am a doctor*"). The test suite from (Dawkins et al., 2024) measures the gender resolution tendencies of MT systems in literary-style dialogues.

However, to the best of our knowledge, none of the available test suites addresses the writer's gender from a general point of view. We believe that the presented test suite will add value to studies addressing gender in the area of machine translation and natural language processing.

3 Test suite creation

The test suite consists of 1 000 user reviews of 10 different product categories extracted from the publicly available repository "Amazon reviews 2023"[2] (Hou et al., 2024). The repository contains reviews written between 1996 and 2023 divided into more than 30 different product categories. For our test suite, we selected 10 categories, and extracted the first 100 reviews from each according to the following criteria:

- take the newest reviews, written in 2023;
- take the reviews with at least 10 occurrences of the first-person pronoun "I";
- take the reviews not longer than 250 (untokenised) words;
- do not include repeated reviews.

The threshold for the pronoun "I" is set to ensure that there will be enough instances of first-person singular gendered words in the translations. The length limit is set to avoid very long reviews, and very short reviews are automatically discarded due to the threshold for the pronoun "I".

The statistics of the obtained test suite is shown in Table 1. The product categories are selected to be distinct, and also to include some stereotypically masculine (e.g. cars, tools and improvements) and feminine (eg. beauty and baby products) as well as supposedly neutral ones (e.g. pet supplies).

In most of the reviews, the gender of the writer is not specified by any information in the English source. Explicit gender cues (e.g. "*I'm a man/woman*" or "*male/female*", "*I was pregnant*") can be found only in 1.5% of the reviews: there are 14 explicitly feminine reviews and one masculine. In addition, some of the reviews contain potential gender cues, namely "*husband*" and "*wife*" which used to be explicit and therefore can influence the choice of the writer's gender. In total, there are 17 reviews with "*my husband*" and 16 reviews with "*my wife*". All in all, there are notably more explicit feminine cues, while potential cues are balanced. Therefore, it can be expected that a translation with balanced gender distribution might contain slightly more feminine reviews. The distribution of the gender cues over product categories can be seen in Appendix A.1.

Validation set The validation set is created in order to check the performance of the evaluation scripts. The set is constructed according to the same rules as the test suite, the only differences are the included years and the size. The reviews were selected from all years before 2023, in order to avoid any overlap with the test suite. From each of the product categories, the first 10 reviews were selected so that it consists of 100 reviews in total, with 1 238 occurrences of the pronoun "I" and 18 789 untokenised running words.

The text is translated into Russian and Serbian using `Gemini 2.5 Flash model`[3] and `Google Translate`[4] in June 2025, and the four translations

[2] https://amazon-reviews-2023.github.io/

[3] https://gemini.google.com/u/1/app

[4] https://translate.google.com

product category	reviews	ocurrences of "I"	untokenised running words
AUTOMOTIVE	10x100	1 207	19 418
BABY PRODUCTS		1 222	19 051
BEAUTY AND PERSONAL CARE		1 229	19 054
HEALTH AND HOUSEHOLD		1 265	18 444
HOME AND KITCHEN		1 261	19 007
MUSICAL INSTRUMENTS		1 209	19 316
PET SUPPLIES		1 216	19 312
SPORTS AND OUTDOORS		1 218	18 562
TOOLS AND HOME IMPROVEMENT		1 207	19 410
VIDEO GAMES		1 286	18 345
total	1 000	12 240	189 919

Table 1: Corpus statistics: English Amazon reviews from ten different product categories: number of reviews, number of occurrences of the first-person pronoun "I", and length (number of untokenised running words).

are used for human assessment of the evaluation scripts described in the following sections.

4 Evaluation method

The evaluation method follows the principles from the manual gender labelling described in (Popovic and Lapshinova-Koltunski, 2024), but is fully automatic. It consists of three steps: (1) word-level annotation (identifying gendered words related to first-person singular), (2) review-level annotation based on the word-level gender labels, and (3) calculation of gender score based on the review-level gender labels.

4.1 Word-level annotation

The word-level annotation consists of identifying and labelling gendered words of interest, namely words referring to the first-person singular. For both languages, the words of interest are verb past participles and adjectives.

The annotation is based on POS tags from Stanza tool (Qi et al., 2020). For each language, a corresponding rule-based Python script is used. Two different scripts are necessary partly due to the differences between languages, and partly because of the differences between the provided POS tags.

Examples of tagged words for each of the languages can be seen in Table 2. For Serbian, both universal POS tags as well as treebank-specific POS tags which contain the information about person, gender, tense, number, etc. are available. For Russian, only universal POS tags are available and the further information can be found only in morpho-syntactic features.

From the given example, it can also be seen that, unfortunately, neither POS tags nor morpho-syntactic features of verb past participles and adjectives contain the information whether they correspond to the first-person singular. Therefore, a span of surrounding words has to also be checked according to the grammatical rules for each language.

In Serbian, words of interest can precede or follow the auxiliary verb "*biti*" (*to be*) in any first-person singular form. Since pronouns are in general more often omitted than not, they cannot be used here. Although the word order is rather free, the distance between the auxiliary verb and a word of interest is usually not larger than 3. Therefore, the context of 3 preceding and 3 following words was included. Increasing the range would serve only for a small number of cases, but would lead to picking up words referring to other persons or objects, thus decreasing the precision and potentially deteriorating the overall performance.

In Russian, words of interest follow the first-person singular personal pronoun я (*I*) and there is no auxiliary verb. In some cases, the pronoun can be placed immediately after the word of interest. Similarly to Serbian, while longer distances are also possible, increasing the span can easily decrease the precision. Therefore, the context of 3 preceding and one following word was included.

The overall process can be described as follows:

- find a potential word of interest (verb past participle or adjective);
- check whether the auxiliary verb/personal pro-

language	word	universal POS	treebank POS	universal morpho-syntactic features
Serbian	to	DET	Pd-nsn	Case=Nom|Gender=Neut|Number=Sing|PronType=Dem
	sam	AUX	**Var1s**	Mood=Ind|Number=Sing|Person=1|Tense=Pres|VerbForm=Fin
	uradio	VERB	**Vmp-sm**	Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part|Voice=Act
Russian	я	PRON	NA	**Case=Nom|Number=Sing|Person=1|PronType=Prs**
	этого	PRON	NA	Animacy=Inan|Case=Gen|Gender=Neut|Number=Sing|PronType=Dem
	делала	VERB	NA	Aspect=Imp|**Gender=Fem**|Mood=Ind|**Number=Sing**|**Tense=Past**|VerbForm=Fin|Voice=Act

Table 2: Examples of words annotated by Stanza tool in Serbian (above) and Russian (below).

noun in first-person singular form can be found in the given context;

- if yes, take the gender of the word of interest and increment the corresponding gender count.

The main difference between the two scripts is related to morpho-syntactic features. While they were immediately available in Serbian tree-bank tags, finding them in Russian required more computational effort. First, the universal POS tag was checked, and then the list of morpho-syntactic features was traversed in order to find additional information about the gender, number, as well as person of surrounding words. For these reasons, the script for Russian is much slower than the one for Serbian.

4.2 Human assessment on the validation set

In order to assess the performance of the word-level annotation, the two `Gemini` translations of the validation set described in Section 3 were annotated by the corresponding scripts. The annotations are then checked by experts, i.e. trained linguists with the native command of the target languages. They were instructed to determine whether the annotated words are correct (precision) as well as whether all gendered words of interest were captured (recall). The results in Table 3 show very high precision (over 99%) for both languages, meaning that almost all annotated words are really referring to the first-person singular. As for recall, it is high for Serbian (95%), but notably smaller for Russian (75%).

A qualitative analysis of errors showed that in both languages, the long-range dependencies were not captured, as expected. Further analysis of the low Russian recall revealed that for a considerable number of frequent adjectives, the relevant information about the nominative case was missing, so that they were not considered as words of interest. If the rule were changed, many other adjectives (referring to other people or objects) would be selected thus decreasing the precision and possibly deteriorating the overall performance.

script validation		
	en-ru	en-sr
precision	99.8	99.7
recall	75.6	95.2

Table 3: Evaluation of annotating scripts on a validation set.

Another problem with Russian is occasional occurrence of informal style where the pronoun is fully omitted. For those cases, it is practically impossible to create a rule for capturing the word of interest because there are no related first-person singular words around.

Because of the low recall for Russian, the review-level annotation (described in the next section) which is essential for the task was manually checked as well, on all four translations of validation set. The resulting scores can be seen in Appendix A.2. Since the review labels were correct and no problems related to word annotation errors were identified, the script is considered well-suited for the task.

It should be noted that the problems with low word-level recall could be addressed by using LLMs. Our initial experiment with LLMs using few-shot prompts was, however, not successful. While being able to increase the recall to some extent, the precision dropped notably because of tagging a large number of irrelevant words (second and third person referring to other people or objects, often even in plural form). A systematic set of experiments with different prompt designs would be necessary, and will be investigated in future work.

4.3 Review-level annotation

For each review, a gender label is assigned according to the gender of the identified words of interest in the previous step. If no words of interest were identified, the review is labelled as "x" (no gender found). Otherwise, a gender proportion score

$$gp = \frac{C(m) - C(f)}{C(m) + C(f)}$$

is calculated, where $C(f)$ denotes the count of feminine words of interest and $C(m)$ presents the masculine count.

$$gp = \begin{cases} \text{feminine}, & gp < -0.4 \\ \text{masculine}, & gp > 0.4 \\ \text{mixed}, & -0.4 \leq gp \leq 0.4 \end{cases}$$

If the $gp < -0.4$, the review is considered feminine, if $gp > 0.4$ masculine. If the score is between -0.4 and 0.4, the review is considered as "mixed". This soft decision approach is chosen to alleviate potential errors of the annotation scripts and also to retain clear tendencies of a translation model towards one gender on the word level.

Table 4 presents an example from the validation set. In the Russian translation, all words of interest are masculine, so that the gender proportion gp is equal to 1 and the review-level gender is masculine. In the Serbian translation, there are five feminine and two masculine words. The gender proportion is then $gp = (2 - 5)/(2 + 5) = -3/7 = -0.43$. Since it is less than -0.4, the review is labelled as feminine. If there were four feminine and two masculine words, the proportion would be $(2 - 4)/(2 + 4) = -2/6 = -.033$, which is between -0.4 and 0.4 so that the review would be labelled as mixed.

4.4 Gender scores

In order to estimate gender bias in a set of reviews, the following score is calculated:

$$genderScore = 100 * \frac{N(m) - N(f)}{N}$$

where $N(m)$ denotes the total number of masculine reviews, $N(f)$ is the total number of feminine reviews, and N is the total number of reviews, including those marked with "x" and the mixed ones. The "x" and mixed reviews are thus not contributing to the score.

It should be explained that the analysis of the "x" reviews goes beyond the scope of this work, so it is not known whether they are really genderless (not containing any words of interest), or written in the gender-neutral way (difficult for both target languages but possible in some cases), or written using some kind of inclusive forms. However, there are not so many systematic consistent strategies for gender-neutral or inclusive forms. They may include the use of plural verb forms with first person singular pronouns or gender-gapping (the use of underscore, e.g. студент_ка (student(m/f)) and also the use of impersonal or indefinite personal structures. It is also common to use both forms, e.g. я был/а разочарован/а or *bio/la sam razočaran/a* (I was(m/f) disappointed(m/f)) in Russian and Serbian, respectively[5]. The POS tagger cannot properly recognise these inclusive forms, and would label them as (proper) nouns.

The values of the gender score range from -100 (all reviews in the text are feminine) to 100 (all reviews are masculine). There are no "good" and "bad" values as such, only the information about the (dis)balance of the two genders in a text. Negative values indicate more feminine reviews, positive values indicate more masculine reviews, and the smaller the absolute value is, the more gender-balanced the text is. We consider the texts with a score between -10 and 10 as gender-balanced.

5 Results on the WMT-2025 translations

5.1 Evaluation levels

In the framework of WMT-2025, the test suite was translated by 40 English→Russian systems and 35 English→Serbian systems. For each language pair, the gender scores are calculated in the following set-ups:

1. language level (all systems together);
2. system level;
3. language level for each product category;
4. system level for each product category.

[5] See more details in (Popovic and Lapshinova-Koltunski, 2024; Popović et al., 2025).

	en	I **came** across an item online with the same concept and **was** completely **interested**. I eventually **bought** them. They were too big for what I **wanted**. I then **bought** these. I **was thinking** of purchasing more.
masc.	ru	Я наткнулся в интернете на предмет с такой же концепцией и полностью заинтересовался. В конце концов я купил их. Они были слишком большими для того, что я хотел. Затем я купил эти. Я думал о покупке еще.
fem.	sr	Naišla sam na predmet na internetu sa istim konceptom i bio sam potpuno zainteresovan. Na kraju sam ih kupila. Bili su preveliki za ono što sam želela. Onda sam kupila ove. Razmišljala sam o kupovini više.

Table 4: Example of gender labels according to first-person singular gendered words.

overall results for each language pair					
		distribution			
lang.	score	m	f	x	mix
en-ru	33.8	25 934	12 427	654	985
en-sr	39.0	23 555	9 920	159	1 366

Table 5: Language level gender scores and label distributions

The gender scores are presented together with the distribution of the review-level gender labels which led to the score value. In the following sections, the results for the first three set-ups are presented and discussed in detail, and the results for the fourth set-up are presented and discussed only for gender-balanced systems. The complete results for all systems and all product categories together with the corresponding discussions are presented in Appendix A.3.

5.2 Language level scores

Overall gender scores for each of the target languages aggregated over all translation outputs are presented in Table 5.

The gender scores are between 30 and 40, indicating that for both languages the majority of the reviews are translated as masculine.

Looking at the distributions of review-level labels, it can be noted that the counts of masculine reviews are similar in both languages, and notably higher than the counts of feminine labels. Furthermore, the number of mixed reviews is notably higher in Serbian, while the number of "x" labels is significantly higher (about 4 times) in Russian. A possible reason for more mixed labels in Serbian is that Serbian is less-resourced than Russian so that there are more translation errors. As for the larger amount of "x" labels in Russian, one possible reason is that Russian systems often generate some kind of inclusive form. Another possibility is the low recall of the annotation script, so that in some of the reviews none of the words of interest are captured. As previously mentioned, the nature of "x" labels was not analysed in this work, but should be part of the future research.

5.3 System level scores

The gender scores and label distributions for each of the participating systems are presented in Tables 6 and 7. The systems are ranked from lowest to highest scores, and, as previously mentioned, the most balanced systems are considered to be those with scores between -10 and 10. It can be noted that for both languages, only seven systems have balanced score values. Five of them, namely Algharb, Gemini-2.5 Pro, Shy, Wenuiil and Yolu are balanced for both languages. GemTrans is among the most balanced for Serbian, but also not very far for Russian with the score of 14.9. ONLINE-G and ONLINE-B have different tendencies in the two languages: balanced for one, but very (46.9) or extremely (86.7) masculine for the other. Moreover, it can be noted that ONLINE-G (in contrast to other balanced systems) generated a high number of mixed genders in both languages. Yandex did not participate in translating into Serbian.

Furthermore, it can be seen that all other systems are masculine-biased, to more or less extent. There is no system with a bias towards feminine writer's gender. Moreover, two systems, TranssionMT and Mistral-7B, are extremely biased for both languages, with (almost) all reviews translated into masculine form – not a single feminine review was identified in TranssionMT outputs.

5.4 Language level scores for different product categories

Table 8 presents the language level gender scores for each of the ten product categories. The tenden-

English→Russian

		distribution			
system	score	m	f	x	mix
Algharb	-0.5	485	490	16	9
Yolu	-1.2	482	494	13	11
Yandex	-1.8	481	499	17	3
Gemini-2.5-Pro	1.7	499	482	16	3
ONLINE-G	2.3	412	389	10	189
Wenyiil	8.3	528	445	19	8
Shy	9.9	533	434	27	6
Laniqo	11.6	537	421	31	11
GemTrans	14.9	568	419	9	4
SalamandraTA	16.8	561	393	15	31
TowerPlus-9B	17.1	579	408	9	4
IRB-MT	18.3	580	397	18	5
hybrid	20.3	580	377	34	9
Claude-4	20.4	590	386	20	4
Gemma-3-12B	23.1	603	372	15	10
GPT-4.1	23.3	607	374	18	1
DeepSeek-V3	24.4	607	363	26	4
ONLINE-W	25.5	528	273	11	188
DLUT_GTCOM	26.5	614	349	20	17
UvA-MT	27.4	630	356	10	4
AyaExpanse-32B	31.7	649	332	11	8
Qwen3-235B	34.4	662	318	12	8
EuroLLM-22B	35.5	665	310	12	13
Gemma-3-27B	38.2	681	299	13	7
CommandA	39.3	686	293	14	7
TowerPlus-72B	40.6	693	287	9	11
TranssionTranslate	44.7	645	198	10	147
ONLINE-B	46.9	715	246	10	29
AyaExpanse-8B	47.6	728	252	9	11
IR-MultiagentMT	47.6	719	243	30	8
Qwen2.5-7B	48.0	681	201	60	58
SRPOL	49.2	733	241	12	14
Llama-4-Maverick	54.0	762	222	13	3
CommandA-MT	54.8	767	219	8	6
CommandR7B	55.0	753	203	17	27
Llama-3.1-8B	55.2	750	198	5	47
EuroLLM-9B	66.7	822	155	17	6
NLLB	83.6	896	60	29	15
Mistral-7B	90.9	938	29	3	30
TranssionMT	98.5	985	0	6	9

Table 6: System level gender scores and review label distributions for Russian

English→Serbian

		distribution			
system	score	m	f	x	mix
Gemini-2.5-Pro	-3.2	483	515	1	1
Algharb	-2.3	485	508	1	6
ONLINE-B	3.7	506	469	1	24
Yolu	4.4	510	466	2	22
GemTrans	6.8	527	459	2	12
Wenyiil	7.0	531	461	1	7
Shy	7.6	535	459	1	5
CUNI-SFT	16.0	549	389	5	57
GPT-4.1	20.7	601	394	1	4
Claude-4	21.4	604	390	1	5
EuroLLM-22B	22.9	589	360	2	49
hybrid	22.9	609	380	3	8
IRB-MT	23.3	608	375	2	15
Gemma-3-12B	27.9	606	327	2	65
AyaExpanse-32B	29.9	622	323	3	52
Gemma-3-27B	32.4	645	321	3	31
UvA-MT	32.5	656	331	1	12
DeepSeek-V3	34.5	668	323	1	8
TowerPlus-9B	36.4	628	264	22	86
AyaExpanse-8B	37.8	626	248	5	121
Qwen3-235B	43.8	715	277	1	7
CommandR7B	45.1	643	192	54	111
Llama-3.1-8B	46.7	707	240	3	50
CommandA	52.4	747	223	1	29
IR-MultiagentMT	52.6	753	227	6	14
EuroLLM-9B	56.9	750	181	2	67
Qwen2.5-7B	58.7	732	145	12	111
SalamandraTA	59.9	765	166	3	66
Llama-4-Maverick	62.5	804	179	2	15
CommandA-MT	69.6	841	145	1	13
TowerPlus-72B	70.8	835	127	3	35
Mistral-7B	86.1	898	37	4	61
ONLINE-G	86.7	878	11	3	108
TranssionMT	90.8	916	8	2	74
TranssionTranslate	98.3	983	0	2	15

Table 7: System level gender scores and review label distributions for Serbian

cies are the same in both languages: feminine bias is present only for two product categories: BEAUTY AND PERSONAL CARE and BABY PRODUCTS. And even though they are clearly "feminine", the gender scores are not lower than -45.

For all other products categories, the majority of the reviews are masculine. The most balanced category is PET SUPPLIES, with the score of 5.5 for Russian and 19.0 for Serbian.

The most masculine products seem to be VIDEO GAMES, MUSICAL INSTRUMENTS and AUTOMOTIVE, with all gender scores over 80. While the biases in BEAUTY AND PERSONAL CARE and BABY PRODUCTS as well as in VIDEO GAMES and AUTOMOTVE are expected due to the widely known stereotypes, the results for MUSICAL INSTRUMENTS are somewhat surprising. The same tendency was already observed in (Popovic and Lapshinova-Koltunski, 2024). However, the results were reported only on a small data set consisting of 14 reviews per category, so they were not reliable. It was nevertheless striking that even the human translators did not opt for the feminine writer's gender for any of the reviews in this category.

As for the rest of the categories, the most masculine one is TOOLS AND HOME IMPROVEMENT with the scores over 60, followed by SPORTS AND OUTDOORS with the scores around 45. The other two, HEALTH AND HOUSEHOLD and HOME AND KITCHEN are more balanced, with scores between 20 and 30.

The label "x" is relatively uniformly distributed over the categories, except BEAUTY AND PERSONAL CARE in Russian, and BABY PRODUCTS in both languages, where a notably higher amount of "x" reviews can be noted. A deeper linguistic analysis for a better understanding of the nature of these translations could, therefore, start from translations of reviews in these product categories.

The amount of mixed reviews is apparently proportional to the amount of feminine reviews. The tendency is confirmed by calculating Pearson correlation coefficients presented in Table 9 (low correlations for the "x" label can be seen, too).

A possible reason might be that the models are intrinsically inclined to choose masculine first-person singular words, so that even when the number of feminine words increases, many of them are still mixed with masculine words within the same review. A deeper analysis of the word-level annotations might reveal more details about the background, and is planned for the future work.

5.5 System level scores for different product categories

The gender scores of each product category for the gender-balanced systems are presented in Table 10. It can be seen that, although the systems are gender-balanced for the entire test suite, they are far from balanced within different product categories. This means that the reason for the overall balance lies in the choice of the product categories and not in the properties of the systems. They are all heavily masculine-biased for each of the three most masculine categories, namely AUTOMOTIVE, MUSICAL INSTRUMENTS and VIDEO GAMES. The overall balance seems to be achieved because this masculine bias is compensated not only by the two most feminine categories BABY PRODUCTS and BEAUTY AND PERSONAL CARE, but also by HEALTH AND HOUSEHOLD, HOME AND KITCHEN and PET SUPPLIES.

As for differences between the languages, ONLINE-G is balanced for Russian but clearly masculine-biased for Serbian for all categories. ONLINE-B is balanced for Serbian, whereas for Russian, HEALTH AND HOUSEHOLD, HOME AND KITCHEN and PET SUPPLIES are predominantly masculine instead of feminine, and the feminine bias for Baby Products is notably smaller.

Other system behave differently for different product categories and no regular patterns were observed, although there are certain tendencies which are discussed in Appendix A.3.

6 Summary and Outlook

Summary The presented test suite is designed for analysis of the first-person singular gender (speaker's or writer's gender) in translations from English into Russian and Serbian. The gender score is defined to measure the balance between the two binary genders, masculine and feminine. The score ranges between -100 (fully feminine) and 100 (fully masculine), and values between -10 and 10 are considered as balanced.

After using the test suite on WMT-2025 translation outputs to calculate language level scores, system level scores and product level scores, the main findings are:

- the majority of the systems are biased towards masculine writer's gender;
- none of the systems is biased towards feminine writer's gender;

			distribution			
product category	lang.	score	m	f	x	mix
Automotive	en-ru	83.0	3610	291	50	49
	en-sr	81.5	3141	287	14	58
Baby Products	en-ru	-28.1	1324	2447	89	140
	en-sr	-14.1	1380	1875	49	196
Beauty and Personal Care	en-ru	-42.7	1033	2742	116	109
	en-sr	-27.6	1155	2122	16	207
Health and Household	en-ru	21.9	2338	1463	67	132
	en-sr	26.9	2130	1189	10	171
Home and Kitchen	en-ru	19.7	2299	1512	58	131
	en-sr	30.3	2192	1132	12	164
Musical Instruments	en-ru	84.5	3648	269	47	36
	en-sr	82.5	3157	270	12	61
Pet Supplies	en-ru	5.5	2000	1780	49	171
	en-sr	19.0	1986	1322	13	179
Sports and Outdoors	en-ru	45.6	2836	1011	43	110
	en-sr	46.7	2494	860	13	133
Tools and Home Improvement	en-ru	63.2	3191	662	68	79
	en-sr	60.9	2755	625	8	112
Video Games	en-ru	85.1	3655	250	67	28
	en-sr	83.6	3165	238	12	85

Table 8: Results for each product category and each language pair

	m	f
mix	-.839	.714
x	-.238	.497

Table 9: Pearson's correlation coefficients between the gender labels in different product categories: the number of mixed reviews is proportional to the number of feminine reviews.

- five systems are gender-balanced for both target languages (with scores between -10 and 10): *Algharb*, *Gemini-2.5-Pro*, *Shy*, *Wenyiil* and *Yolu*;
- one model (*GemTrans*) is only slightly unbalanced towards masculine for Russian (score 14.9)
- two models behave differently depending on the language (*ONLINE-B* and *ONLINE-G*)
- one model (*Yandex*) participated only for Russian

Further analysis of different product categories showed that none of the systems is balanced within all product categories, while some systems are balanced within one single product category. This means that the overall gender balance is a consequence of the choice of product categories, not of the system properties.

Furthermore, three product categories are identified to be predominantly masculine and two as mostly feminine. Automotive, Musical Instruments and Video Games are heavily biased towards masculine by all models (all scores are larger than 50). Baby Products and Beauty and Personal Care are biased towards feminine, but to a lesser extent: scores are ranging from -80 to 99, and a few systems are balanced.

Overall, the results confirmed the similar tendencies reported in previous work for Croatian and Russian on a small scale (Popovic and Lapshinova-Koltunski, 2024). Given that one of the reported tendencies was that even human translators are generally inclined to opt more often for a masculine writer, and are also influenced by the product category, the WMT-2025 results from the test suite are not surprising.

Outlook The test suite offers several possibilities for future work, some of them are already mentioned in previous sections. One important open

system		all	Auto	Baby	Beauty	Health	Home	Music	Pets	Sports	Tools	Games
Algharb	ru	-0.5	75	-79	-74	-36	-29	74	-61	**8**	38	79
	sr	-2.3	72	-83	-79	-32	-35	77	-61	**6**	34	78
Gemini	ru	1.7	71	-77	-73	-31	-22	74	-52	17	33	77
	sr	-3.2	74	-83	-77	-36	-32	72	-66	**4**	32	80
GemTrans	ru	*14.9*	85	-64	-74	**-9**	-17	82	-37	35	58	90
	sr	6.8	74	-60	-81	-30	-17	74	-39	28	38	81
ONLINE-B	ru	*46.9*	*93*	*-24*	*-62*	*52*	*40*	*95*	*38*	*62*	*77*	*98*
	sr	3.7	77	-72	-84	-43	-26	80	-47	31	33	88
ONLINE-G	ru	2.3	58	-43	-79	-53	-49	76	-29	24	39	79
	sr	*86.7*	*97*	*63*	*74*	*88*	*82*	*98*	*89*	*89*	*91*	*96*
Shy	ru	9.9	76	-65	-69	-15	**-10**	76	-41	22	50	75
	sr	7.6	72	-73	-74	-23	-12	78	-43	27	46	78
Wenyiil	ru	8.3	81	-75	-71	-20	-16	75	-39	17	48	83
	sr	7.0	74	-74	-74	-15	-12	86	-47	**10**	44	78
Yolu	ru	-1.2	71	-67	-73	-30	-45	78	-50	12	29	63
	sr	4.4	72	-65	-86	-22	-26	70	-39	22	41	77
Yandex	ru	-1.8	73	-73	-75	-39	-31	72	-63	22	36	60

Table 10: Gender scores for the most balanced systems: overall, and for each product category. Balanced product category scores are presented in bold.

question is the nature of the translations with the label "x". It should be explored whether they are really gender-neutral, or an inclusive form was used, or the annotation script missed all words of interest, or there are possibly translation errors. As for mixed reviews, word-level gender scores could be added to include their contribution. Another direction is analysis of reviews with gender cues (for instance specific words like *pregnant*, etc., and their possible separation into a sub-suite.

An obvious direction is extending the test suite with more data. Also, it can be easily extended to other languages with gendered first-person singular words, such as French, Spanish, Czech among others. Also, other domains/genres apart from Amazon product reviews should be considered.

Finally, a systematic experiment on prompt design should be carried out in order to use LLMs for word-level annotation and improve recall without deteriorating precision.

Limitations

The presented test suite comes with a few limitations. Currently, it deals only with two target languages, both of them being Slavic although with different grammar rules. Furthermore, only one evaluation method has been systematically explored so far, namely using Stanza POS tags for a rule-based identification of gendered words. The script performs well for Serbian, but has notably lower recall for Russian, and also a high time complexity, due to the differences between the POS tags provided by the tool. Moreover, it should be kept in mind that mixed and "x" reviews were excluded from the gender score, but not analysed. Also, the reviews with explicit or implicit gender cues are not analysed.

References

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181, Florence, Italy. Association for Computational Linguistics.

Hillary Dawkins, Isar Nejadgholi, and Chi-kiu Lo. 2024. WMT24 test suite: Gender resolution in speaker-listener dialogue roles. In *Proceedings of the Ninth Conference on Machine Translation*, pages 307–326, Miami, Florida, USA. Association for Computational Linguistics.

Nizar Habash, Houda Bouamor, and Christine Chung. 2019. Automatic gender identification and reinflection in Arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 155–165, Florence, Italy. Association for Computational Linguistics.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.

Ekaterina Lapshinova-Koltunski, Maja Popović, and Maarit Koponen. 2022. DiHuTra: a parallel corpus to analyse differences between human translations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1751–1760, Marseille, France. European Language Resources Association.

Maja Popovic and Ekaterina Lapshinova-Koltunski. 2024. Gender and bias in Amazon review translations: by humans, MT systems and ChatGPT. In *Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies*, pages 22–30, Sheffield, United Kingdom. European Association for Machine Translation (EAMT).

Maja Popović, Ekaterina Lapshinova-Koltunski, and Anastasiia Göldner. 2025. Did I (she) or I (he) buy this? or rather I (she/he)? towards first-person gender neutral translation by LLMs. In *Proceedings of the 3rd Workshop on Gender-Inclusive Translation Technologies (GITT 2025)*, pages 64–73, Geneva, Switzerland. European Association for Machine Translation.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Beatrice Savoldi, Marco Gaido, Matteo Negri, and Luisa Bentivogli. 2023. Test suites task: Evaluation of gender fairness in mt with MuST-SHE and INES. In *Proceedings of the Eighth Conference on Machine Translation*, pages 252–262, Singapore. Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Matteo Negri, and Luisa Bentivogli. 2024. FBK@IWSLT test suites task: Gender bias evaluation with MuST-SHE. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 65–71, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

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.

Eva Vanmassenhove and Johanna Monti. 2021. gENder-IT: An annotated English-Italian parallel challenge set for cross-linguistic natural gender phenomena. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 1–7, Online. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

A Appendix

A.1 Explicit and potential gender cues

Table 11 presents the distribution of explicit ("*I am a man/woman*" or "*male/female*", "*I am pregnant*", etc.) and potential ("*my husband*", "*my wife*") gender cues over different product categories, mentioned in Section 3.

The single explicit masculine cue occurs in the AUTOMOTIVE category. The explicit feminine cues are distributed over several categories, most frequently in BABY PRODUCTS and TOOLS AND HOME IMPROVEMENT followed by BEAUTY AND PERSONAL CARE and SPORTS AND OUTDOORS.

A.2 Review labels on the validation set

Table 12 presents gender scores and label distributions for two translations of the validation set described in Section 3. The evaluation scripts were run on the two translations of the validation set, the one generated by Gemini (which was also used for assessing word-level annotation), and another one generated by Google Translate. The review labels were checked by human annotators in order to assess the influence of word-level errors on the review-level labels (and thus on the final score).

A.3 Product categories: system level

This section presents and discusses the system level scores for each product category (Tables 13–32), as mentioned in Section 5. While systems generally behave differently for different products, certain tendencies can be observed in each of the product categories. The overall gender balanced systems are presented in italic. It is already discussed in Section5 that they are far from balanced for most of the categories, being clearly biased towards one or other writer's gender. In this section it can be seen that they are less biased in the three heavily masculine categories than other systems, and more biased in the two heavily feminine categories than other systems, but also clearly feminine biased in categories with relatively uniform distribution of feminine and masculine scores (e.g. HEALTH AND HOUSEHOLD, HOME AND KITCHEN and PET SUPPLIES.

	feminine cues		masculine cues	
	explicit	potential	explicit	potential
Automotive	0	1	1	3
Baby Products	3	1	0	0
Beauty and Personal Care	2	2	0	0
Health and Household	1	0	0	2
Home and Kitchen	1	4	0	0
Musical Instruments	1	2	0	4
Pet Supplies	0	2	0	0
Sports and Outdoors	2	4	0	2
Tools and Home Improvement	3	0	0	1
Video Games	0	1	0	3
total	14	17	1	16

Table 11: Number of reviews (in each product category and overall) with explicit (man, woman, pregnancy, female, male) and potential (husband, wife) gender cues.

			distribution			
	system	score	m	f	x	mix
en-ru	Gemini	-17.1	40	57	3	0
	Google	48.0	71	23	3	3
en-sr	Gemini	-20.0	39	59	1	1
	Google	4.0	48	44	3	5

Table 12: Gender scores and label distributions for two translations of the validation set

Automotive (Tables 13 and 14) This category is overall heavily masculine-biased, with all system scores over 50. Even the overall gender balanced systems are clearly masculine for these products, with scores between 70 and 80.

Baby Products (Tables 15 and 16) The majority of gender scores is feminine-biased, however, there are a few notably masculine-biased outputs. There are 6 balanced systems for Russian and 4 for Serbian, and none of them is balanced for both languages. As for overall gender balanced systems, all of them are clearly feminine-biased with the scores ranging from -83 to -43.

Beauty and Personal Care (Tables 17 and 18) The majority of gender scores is feminine-biased, however, there are a few notably masculine-biased outputs. There are 2 balanced systems for Russian and 5 for Serbian. `CommandR7B` is balanced for both languages, more inclined to feminine for Russian and to masculine in Serbian. Interestingly, `Llama-3.1-8B` for Serbian is perfectly gender-balanced with the score equal to 0, though with 9 mixed and one "x" review. All overall gender balanced systems are clearly feminine-biased with the scores ranging from -86 to -69.

Health and Household (Tables 19 and 20) There are both feminine and masculine gender scores, however more systems are masculine-biased. There are 6 balanced systems for Russian and 2 for Serbian. `Gemma-3-12B` is balanced for both languages, more inclined to feminine for Russian and to masculine in Serbian. All overall gender balanced systems are feminine-biased, although to a less extent than for the previous two categories, with the scores ranging from -53 to -12.

Home and Kitchen (Tables 21 and 22) There are both feminine and masculine gender scores, however more systems are masculine-biased. There are 10 balanced systems for Russian and 3 for Serbian. *Claude-4* and `GPT-4.1` are balanced for both languages. `GPT-4.1` is more inclined to masculine for Russian and to feminine in Serbian. `Claude-4` is inclined to feminine in Russian and perfectly balanced in Serbian, with the score of 0 and no mixed or "x" reviews. Another perfectly balanced system is `DLUT_GTCOM` for Russian, however with 2 mixed and 2 "x" reviews. All overall gender balanced systems are feminine-biased, although to a less extent than for the previous three categories, with the scores ranging from -49 to -10 (`Shy` for Russian is considered as balanced with the score -10 exactly on the threshold).

Musical Instruments (Tables 23 and 24) This category is overall heavily masculine-biased, with all system scores over 50. Even the overall gender-balanced systems are heavily masculine-biased with the scores between 70 and 86.

Pet Supplies (Tables 25 and 26) There are both feminine and masculine gender scores, relatively balanced proportion of systems in Russian however more masculine systems in Serbian. There are 6 balanced systems for Russian and 3 for Serbian. `AyaExpanse-32B` is balanced for both languages, more inclined to feminine for Russian and to masculine in Serbian. All overall gender balanced systems are feminine-biased, with the scores ranging from -66 to -29.

Sports and Outdoors (Tables 27 and 28) There are no feminine gender scores, and there are two balanced systems (inclined to masculine) for each language. For Serbian, both those systems, `Algharb` and `Gemini-2.5-Pro`, are also balanced overall, while for Russian it is only `Algharb`. Other overall balanced systems are clearly although not heavily masculine, with the scores ranging from 12 to 31.

Tools and Home Improvement (Tables 29 and 30) There are no feminine gender scores, and no balanced systems: the scores range from 29 to 100. Although this category is not so heavily masculine-biased as others, even the overall gender balanced systems are masculine-biased for this one with the scores between 29 and 50.

Video Games (Tables 31 and 32) This category is overall heavily masculine-biased, with all system scores over 50. Even the overall gender-balanced systems are very masculine-biased with the scores ranging between 60 and 88.

AUTOMOTIVE, English→Russian

		distribution			
system	score	m	f	x	mix
SalamandraTA	57.0	77	20	1	2
ONLINE-G	*58.0*	69	11	0	20
Laniqo	64.0	82	18	0	0
Gemini-2.5-Pro	*71.0*	85	14	1	0
Yolu	*71.0*	85	14	0	1
Yandex	*73.0*	86	13	1	0
TowerPlus-9B	74.0	86	12	1	1
Algharb	*75.0*	86	11	1	2
CommandR7B	76.0	86	10	2	2
Shy	*76.0*	87	11	2	0
UvA-MT	81.0	90	9	1	0
Wenyiil	*81.0*	90	9	1	0
DeepSeek-V3	82.0	90	8	2	0
IRB-MT	82.0	89	7	4	0
Qwen3-235B	82.0	90	8	1	1
AyaExpanse-32B	83.0	91	8	0	1
Gemma-3-12B	83.0	90	7	2	1
Qwen2.5-7B	83.0	88	5	5	2
Claude-4	84.0	91	7	2	0
AyaExpanse-8B	85.0	92	7	1	0
GemTrans	85.0	92	7	1	0
SRPOL	85.0	92	7	1	0
GPT-4.1	86.0	92	6	2	0
Llama-3.1-8B	86.0	92	6	1	1
TranssionTranslate	86.0	89	3	1	7
TowerPlus-72B	87.0	93	6	1	0
DLUT_GTCOM	88.0	93	5	2	0
hybrid	88.0	93	5	2	0
CommandA-MT	89.0	94	5	1	0
Gemma-3-27B	89.0	94	5	1	0
ONLINE-W	89.0	92	3	0	5
EuroLLM-22B	90.0	95	5	0	0
Llama-4-Maverick	90.0	94	4	2	0
IR-MultiagentMT	91.0	94	3	2	1
EuroLLM-9B	92.0	95	3	2	0
CommandA	93.0	96	3	1	0
ONLINE-B	93.0	96	3	1	0
NLLB	95.0	97	2	1	0
Mistral-7B	97.0	98	1	0	1
TranssionMT	99.0	99	0	0	1

Table 13: AUTOMOTIVE, Russian

AUTOMOTIVE, English→Serbian

		distribution			
system	score	m	f	x	mix
CUNI-SFT	59.0	79	20	1	0
AyaExpanse-32B	66.0	82	16	0	2
CommandR7B	70.0	80	10	6	4
EuroLLM-22B	70.0	83	13	0	4
Algharb	*72.0*	86	14	0	0
Shy	*72.0*	86	14	0	0
Yolu	*72.0*	84	12	1	3
Gemini-2.5-Pro	*74.0*	87	13	0	0
GemTrans	*74.0*	87	13	0	0
UvA-MT	74.0	87	13	0	0
Wenyiil	*74.0*	87	13	0	0
ONLINE-B	*77.0*	88	11	0	1
AyaExpanse-8B	78.0	86	8	0	6
Claude-4	78.0	89	11	0	0
EuroLLM-9B	78.0	87	9	0	4
IRB-MT	78.0	89	11	0	0
Gemma-3-12B	79.0	87	8	0	5
TowerPlus-9B	80.0	87	7	1	5
Qwen2.5-7B	81.0	86	5	2	7
hybrid	82.0	91	9	0	0
GPT-4.1	84.0	92	8	0	0
Llama-3.1-8B	86.0	92	6	0	2
DeepSeek-V3	87.0	93	6	0	1
Gemma-3-27B	87.0	93	6	0	1
SalamandraTA	87.0	92	5	0	3
Qwen3-235B	88.0	94	6	0	0
IR-MultiagentMT	89.0	93	4	2	1
CommandA	90.0	95	5	0	0
TowerPlus-72B	91.0	95	4	0	1
CommandA-MT	92.0	96	4	0	0
Mistral-7B	94.0	96	2	1	1
ONLINE-G	97.0	97	0	0	3
TranssionMT	97.0	97	0	0	3
Llama-4-Maverick	98.0	99	1	0	0
TranssionTranslate	99.0	99	0	0	1

Table 14: AUTOMOTIVE, Serbian

BABY PRODUCTS, English→Russian

		distribution			
system	score	m	f	x	mix
Algharb	*-79.0*	9	88	2	1
Gemini-2.5-Pro	*-77.0*	10	87	2	1
Wenyiil	*-75.0*	10	85	3	2
Yandex	*-73.0*	12	85	3	0
Yolu	*-67.0*	15	82	2	1
Shy	*-65.0*	16	81	3	0
GemTrans	-64.0	17	81	2	0
hybrid	-62.0	17	79	3	1
Claude-4	-58.0	20	78	1	1
GPT-4.1	-58.0	19	77	3	1
IRB-MT	-56.0	21	77	1	1
DeepSeek-V3	-54.0	21	75	3	1
Laniqo	-51.0	22	73	4	1
Gemma-3-12B	-47.0	25	72	2	1
UvA-MT	-46.0	26	72	1	1
DLUT_GTCOM	-44.0	24	68	3	5
ONLINE-G	-43.0	17	60	1	22
Gemma-3-27B	-40.0	28	68	3	1
SalamandraTA	-39.0	28	67	3	2
AyaExpanse-32B	-38.0	29	67	4	0
CommandA	-36.0	30	66	3	1
TowerPlus-9B	-36.0	31	67	2	0
ONLINE-W	-34.0	17	51	2	30
Qwen3-235B	-30.0	33	63	3	1
EuroLLM-22B	-29.0	34	63	2	1
ONLINE-B	-24.0	36	60	1	3
TowerPlus-72B	-23.0	36	59	3	2
SRPOL	-17.0	39	56	2	3
Qwen2.5-7B	-14.0	38	52	3	7
IR-MultiagentMT	-10.0	43	53	3	1
TranssionTranslate	-10.0	32	42	1	25
CommandA-MT	-6.0	46	52	1	1
AyaExpanse-8B	-4.0	46	50	2	2
Llama-4-Maverick	-4.0	47	51	2	0
Llama-3.1-8B	3.0	48	45	1	6
CommandR7B	14.0	54	40	2	4
EuroLLM-9B	22.0	59	37	3	1
NLLB	74.0	85	11	2	2
Mistral-7B	80.0	87	7	0	6
TranssionMT	97.0	97	0	2	1

Table 15: BABY PRODUCTS, Russian

BABY PRODUCTS, English→Serbian

		distribution			
system	score	m	f	x	mix
Algharb	*-83.0*	8	91	1	0
Gemini-2.5-Pro	*-83.0*	8	91	1	0
Wenyiil	*-74.0*	12	86	1	1
Shy	*-73.0*	13	86	1	0
ONLINE-B	*-72.0*	13	85	1	1
Yolu	*-65.0*	17	82	0	1
GemTrans	*-60.0*	19	79	1	1
hybrid	-57.0	20	77	1	2
Claude-4	-55.0	22	77	1	0
GPT-4.1	-52.0	23	75	1	1
IRB-MT	-51.0	22	73	1	4
Gemma-3-27B	-44.0	27	71	1	1
CUNI-SFT	-39.0	28	67	2	3
Gemma-3-12B	-39.0	26	65	1	8
EuroLLM-22B	-33.0	31	64	1	4
UvA-MT	-33.0	32	65	1	2
DeepSeek-V3	-23.0	37	60	1	2
AyaExpanse-32B	-20.0	36	56	1	7
Llama-3.1-8B	-13.0	39	52	2	7
Qwen3-235B	-10.0	44	54	1	1
TowerPlus-9B	-9.0	39	48	3	10
IR-MultiagentMT	8.0	52	44	1	3
CommandA	9.0	51	42	1	6
Qwen2.5-7B	11.0	44	33	4	19
AyaExpanse-8B	12.0	47	35	0	18
Llama-4-Maverick	16.0	56	40	2	2
SalamandraTA	19.0	52	33	2	13
CommandA-MT	25.0	60	35	1	4
CommandR7B	25.0	54	29	7	10
EuroLLM-9B	26.0	61	35	1	3
TowerPlus-72B	32.0	62	30	0	8
ONLINE-G	63.0	70	7	2	21
Mistral-7B	70.0	77	7	1	15
TranssionMT	82.0	83	1	1	15
TranssionTranslate	95.0	95	0	2	3

Table 16: BABY PRODUCTS, Serbian

BEAUTY, English→Russian

		distribution			
system	score	m	f	x	mix
ONLINE-G	*-79.0*	4	83	2	11
ONLINE-W	-78.0	7	85	1	7
Yandex	*-75.0*	11	86	3	0
Algharb	*-74.0*	10	84	6	0
GemTrans	-74.0	11	85	2	2
Claude-4	-73.0	11	84	5	0
Gemini-2.5-Pro	*-73.0*	11	84	4	1
Yolu	*-73.0*	11	84	3	2
hybrid	-71.0	12	83	5	0
Wenyiil	*-71.0*	11	82	4	3
Shy	*-69.0*	13	82	5	0
IRB-MT	-67.0	14	81	3	2
Gemma-3-12B	-66.0	15	81	2	2
GPT-4.1	-64.0	16	80	4	0
DeepSeek-V3	-62.0	16	78	5	1
ONLINE-B	-62.0	16	78	1	5
DLUT_GTCOM	-60.0	19	79	2	0
Laniqo	-60.0	16	76	7	1
UvA-MT	-59.0	19	78	1	2
Gemma-3-27B	-58.0	19	77	2	2
TowerPlus-9B	-53.0	22	75	2	1
CommandA	-51.0	22	73	2	3
CommandA-MT	-50.0	24	74	1	1
Qwen3-235B	-49.0	24	73	2	1
TranssionTranslate	-49.0	16	65	1	18
EuroLLM-22B	-48.0	24	72	2	2
AyaExpanse-32B	-45.0	26	71	1	2
SalamandraTA	-45.0	26	71	1	2
TowerPlus-72B	-44.0	26	70	1	3
IR-MultiagentMT	-34.0	31	65	3	1
Llama-4-Maverick	-34.0	31	65	2	2
AyaExpanse-8B	-26.0	35	61	3	1
SRPOL	-19.0	38	57	3	2
Qwen2.5-7B	-18.0	32	50	8	10
Llama-3.1-8B	-13.0	38	51	2	9
CommandR7B	-10.0	42	52	1	5
EuroLLM-9B	-2.0	47	49	3	1
NLLB	70.0	81	11	8	0
Mistral-7B	80.0	87	7	2	4
TranssionMT	99.0	99	0	1	0

Table 17: BEAUTY AND PERSONAL CARE, Russian

BEAUTY, English→Serbian

		distribution			
system	score	m	f	x	mix
Yolu	*-86.0*	6	92	0	2
ONLINE-B	*-84.0*	6	90	0	4
GemTrans	*-81.0*	9	90	1	0
Algharb	*-79.0*	10	89	0	1
Gemini-2.5-Pro	*-77.0*	11	88	0	1
GPT-4.1	-74.0	13	87	0	0
Shy	*-74.0*	12	86	0	2
Wenyiil	*-74.0*	11	85	0	4
hybrid	-72.0	14	86	0	0
IRB-MT	-69.0	15	84	1	0
Gemma-3-12B	-68.0	15	83	0	2
Claude-4	-66.0	16	82	0	2
UvA-MT	-60.0	19	79	0	2
DeepSeek-V3	-58.0	20	78	0	2
EuroLLM-22B	-53.0	21	74	1	4
Gemma-3-27B	-53.0	22	75	0	3
CUNI-SFT	-41.0	26	67	0	7
Qwen3-235B	-41.0	29	70	0	1
AyaExpanse-32B	-28.0	33	61	0	6
IR-MultiagentMT	-27.0	35	62	0	3
CommandA	-25.0	35	60	0	5
TowerPlus-9B	-21.0	33	54	3	10
EuroLLM-9B	-19.0	37	56	0	7
AyaExpanse-8B	-18.0	30	48	1	21
CommandA-MT	-7.0	45	52	0	3
Llama-4-Maverick	-4.0	46	50	0	4
Llama-3.1-8B	**0.0**	45	45	1	9
SalamandraTA	2.0	47	45	0	8
CommandR7B	3.0	41	38	6	15
Qwen2.5-7B	25.0	55	30	1	14
TowerPlus-72B	40.0	66	26	0	8
Mistral-7B	72.0	80	8	1	11
ONLINE-G	74.0	75	1	0	24
TranssionMT	85.0	86	1	0	13
TranssionTranslate	91.0	91	0	0	9

Table 18: BEAUTY AND PERSONAL CARE, Serbian

HEALTH, English→Russian

system	score	distribution			
		m	f	x	mix
ONLINE-G	*-53.0*	11	64	1	24
Yandex	*-39.0*	29	68	2	1
Algharb	*-36.0*	30	66	1	3
Gemini-2.5-Pro	-31.0	34	65	1	0
Yolu	*-30.0*	34	64	1	1
Laniqo	-26.0	34	60	4	2
ONLINE-W	-25.0	23	48	1	28
Wenyiil	*-20.0*	39	59	2	0
Shy	*-15.0*	41	56	1	2
IRB-MT	-14.0	41	55	2	2
GemTrans	-9.0	44	53	2	1
DLUT_GTCOM	-4.0	46	50	1	3
Gemma-3-12B	-1.0	48	49	1	2
DeepSeek-V3	3.0	49	46	4	1
SalamandraTA	5.0	49	44	2	5
TowerPlus-9B	5.0	52	47	1	0
Claude-4	14.0	56	42	2	0
hybrid	14.0	54	40	4	2
UvA-MT	15.0	57	42	1	0
EuroLLM-22B	19.0	58	39	2	1
AyaExpanse-32B	23.0	60	37	1	2
TranssionTranslate	27.0	54	27	1	18
Qwen3-235B	28.0	63	35	1	1
GPT-4.1	29.0	64	35	1	0
CommandA	35.0	66	31	1	2
Qwen2.5-7B	44.0	66	22	7	5
Gemma-3-27B	49.0	73	24	1	2
Llama-3.1-8B	49.0	72	23	0	5
TowerPlus-72B	50.0	74	24	1	1
AyaExpanse-8B	51.0	73	22	1	4
SRPOL	51.0	75	24	1	0
ONLINE-B	52.0	74	22	1	3
IR-MultiagentMT	53.0	73	20	6	1
CommandR7B	61.0	79	18	1	2
Llama-4-Maverick	69.0	84	15	1	0
CommandA-MT	72.0	85	13	1	1
EuroLLM-9B	84.0	91	7	1	1
NLLB	87.0	91	4	3	2
Mistral-7B	91.0	94	3	0	3
TranssionMT	98.0	98	0	1	1

Table 19: HEALTH AND HOUSEHOLD, Russian

HEALTH, English→Serbian

system	score	distribution			
		m	f	x	mix
ONLINE-B	*-43.0*	27	70	0	3
Gemini-2.5-Pro	*-36.0*	32	68	0	0
Algharb	*-32.0*	33	65	0	2
GemTrans	*-30.0*	34	64	0	2
Shy	*-23.0*	38	61	0	1
Yolu	*-22.0*	38	60	0	2
Wenyiil	*-15.0*	42	57	0	1
IRB-MT	-12.0	43	55	0	2
CUNI-SFT	-5.0	44	49	0	7
Gemma-3-12B	8.0	50	42	1	7
DeepSeek-V3	11.0	55	44	0	1
AyaExpanse-8B	13.0	48	35	0	17
Claude-4	13.0	56	43	0	1
hybrid	14.0	56	42	0	2
EuroLLM-22B	15.0	54	39	0	7
GPT-4.1	15.0	57	42	0	1
UvA-MT	19.0	59	40	0	1
Gemma-3-27B	20.0	57	37	0	6
AyaExpanse-32B	23.0	59	36	0	5
Llama-3.1-8B	31.0	61	30	0	9
Qwen3-235B	33.0	66	33	0	1
TowerPlus-9B	33.0	60	27	2	11
CommandR7B	34.0	58	24	6	12
CommandA	50.0	74	24	0	2
SalamandraTA	58.0	73	15	0	12
IR-MultiagentMT	59.0	79	20	0	1
EuroLLM-9B	60.0	75	15	0	10
Llama-4-Maverick	64.0	81	17	0	2
Qwen2.5-7B	70.0	80	10	1	9
TowerPlus-72B	75.0	85	10	0	5
CommandA-MT	84.0	92	8	0	0
Mistral-7B	85.0	89	4	0	7
TranssionMT	85.0	87	2	0	11
ONLINE-G	88.0	89	1	0	10
TranssionTranslate	99.0	99	0	0	1

Table 20: HEALTH AND HOUSEHOLD, Serbian

HOME AND KITCHEN, English→Russian

		distribution			
system	score	m	f	x	mix
ONLINE-G	*-49.0*	15	64	2	19
Yolu	*-45.0*	27	72	1	0
Yandex	*-31.0*	34	65	1	0
Algharb	*-29.0*	34	63	1	2
Gemini-2.5-Pro	*-22.0*	38	60	1	1
GemTrans	-17.0	41	58	1	0
Wenyiil	*-16.0*	41	57	1	1
ONLINE-W	-11.0	29	40	2	29
Shy	*-10.0*	43	53	2	2
IRB-MT	-5.0	47	52	1	0
TowerPlus-9B	-5.0	47	52	1	0
Gemma-3-12B	-3.0	48	51	1	0
Claude-4	-2.0	48	50	1	1
Laniqo	-2.0	47	49	2	2
DLUT_GTCOM	**0.0**	48	48	2	2
hybrid	1.0	50	49	1	0
UvA-MT	4.0	51	47	2	0
GPT-4.1	7.0	53	46	1	0
SalamandraTA	13.0	55	42	1	2
DeepSeek-V3	19.0	59	40	1	0
AyaExpanse-32B	22.0	60	38	1	1
EuroLLM-22B	24.0	60	36	1	3
Gemma-3-27B	24.0	61	37	1	1
Qwen3-235B	29.0	63	34	1	2
TranssionTranslate	31.0	53	22	2	23
AyaExpanse-8B	32.0	65	33	1	1
TowerPlus-72B	34.0	66	32	1	1
ONLINE-B	40.0	66	26	1	7
IR-MultiagentMT	43.0	69	26	4	1
CommandA	44.0	71	27	1	1
Qwen2.5-7B	46.0	61	15	11	13
SRPOL	49.0	74	25	1	0
CommandR7B	53.0	74	21	1	4
Llama-3.1-8B	58.0	76	18	0	6
CommandA-MT	59.0	79	20	1	0
Llama-4-Maverick	62.0	80	18	1	1
EuroLLM-9B	70.0	84	14	1	1
NLLB	84.0	91	7	2	0
Mistral-7B	87.0	92	5	0	3
TranssionMT	99.0	99	0	0	1

Table 21: HOME AND KITCHEN, Russian

HOME AND KITCHEN, English→Serbian

		distribution			
system	score	m	f	x	mix
Algharb	*-35.0*	32	67	0	1
Gemini-2.5-Pro	*-32.0*	34	66	0	0
ONLINE-B	*-26.0*	34	60	0	6
Yolu	*-26.0*	36	62	1	1
GemTrans	*-17.0*	40	57	0	3
Shy	*-12.0*	44	56	0	0
Wenyiil	*-12.0*	44	56	0	0
GPT-4.1	-6.0	47	53	0	0
CUNI-SFT	-5.0	44	49	0	7
Claude-4	**0.0**	50	50	**0**	**0**
hybrid	14.0	56	42	1	1
EuroLLM-22B	15.0	55	40	0	5
IRB-MT	18.0	58	40	0	2
Gemma-3-12B	24.0	53	29	0	18
TowerPlus-9B	27.0	56	29	5	10
AyaExpanse-32B	29.0	63	34	0	3
Gemma-3-27B	30.0	61	31	1	7
Qwen3-235B	32.0	66	34	0	0
UvA-MT	32.0	65	33	0	2
AyaExpanse-8B	35.0	62	27	0	11
DeepSeek-V3	36.0	68	32	0	0
Llama-3.1-8B	39.0	67	28	0	5
CommandR7B	41.0	60	19	4	17
IR-MultiagentMT	49.0	74	25	0	1
Qwen2.5-7B	51.0	70	19	0	11
CommandA	54.0	76	22	0	2
EuroLLM-9B	59.0	77	18	0	5
SalamandraTA	62.0	77	15	0	8
Llama-4-Maverick	69.0	84	15	0	1
CommandA-MT	75.0	86	11	0	3
TowerPlus-72B	77.0	87	10	0	3
ONLINE-G	82.0	82	0	0	18
Mistral-7B	88.0	91	3	0	6
TranssionMT	94.0	94	0	0	6
TranssionTranslate	99.0	99	0	0	1

Table 22: HOME AND KITCHEN, Serbian

MUSICAL INSTRUMENTS, English→Russian

		distribution			
system	score	m	f	x	mix
SalamandraTA	69.0	83	14	1	2
Laniqo	70.0	83	13	4	0
TowerPlus-9B	70.0	85	15	0	0
Yandex	*72.0*	86	14	0	0
Algharb	*74.0*	87	13	0	0
Gemini-2.5-Pro	*74.0*	86	12	2	0
Wenyiil	*75.0*	87	12	1	0
ONLINE-G	*76.0*	82	6	1	11
Shy	*76.0*	87	11	2	0
DeepSeek-V3	77.0	86	9	4	1
Yolu	*78.0*	88	10	1	1
hybrid	80.0	88	8	4	0
Claude-4	81.0	90	9	1	0
IRB-MT	81.0	90	9	1	0
AyaExpanse-32B	82.0	91	9	0	0
AyaExpanse-8B	82.0	91	9	0	0
GemTrans	82.0	91	9	0	0
GPT-4.1	83.0	91	8	1	0
CommandA	85.0	92	7	1	0
Gemma-3-12B	85.0	92	7	1	0
DLUT_GTCOM	87.0	92	5	3	0
Qwen3-235B	87.0	93	6	1	0
SRPOL	87.0	93	6	1	0
Llama-4-Maverick	88.0	94	6	0	0
TowerPlus-72B	88.0	94	6	0	0
CommandR7B	89.0	94	5	1	0
NLLB	89.0	91	2	6	1
EuroLLM-22B	90.0	95	5	0	0
UvA-MT	90.0	95	5	0	0
IR-MultiagentMT	91.0	95	4	0	1
Qwen2.5-7B	91.0	93	2	2	3
CommandA-MT	92.0	96	4	0	0
EuroLLM-9B	93.0	96	3	1	0
Gemma-3-27B	93.0	96	3	1	0
ONLINE-W	93.0	93	0	1	6
Llama-3.1-8B	94.0	95	1	1	3
TranssionTranslate	94.0	94	0	1	5
ONLINE-B	95.0	96	1	1	2
Mistral-7B	97.0	98	1	1	0
TranssionMT	99.0	99	0	1	0

Table 23: MUSICAL INSTRUMENTS, Russian

MUSICAL INSTRUMENTS, English→Serbian

		distribution			
system	score	m	f	x	mix
CUNI-SFT	59.0	77	18	0	5
AyaExpanse-8B	69.0	81	12	1	6
AyaExpanse-32B	70.0	84	14	0	2
Yolu	*70.0*	84	14	0	2
CommandR7B	72.0	79	7	10	4
Gemini-2.5-Pro	*72.0*	86	14	0	0
TowerPlus-9B	72.0	83	11	0	6
EuroLLM-22B	73.0	84	11	0	5
GemTrans	*74.0*	87	13	0	0
Algharb	*77.0*	88	11	0	1
Claude-4	78.0	89	11	0	0
Shy	*78.0*	89	11	0	0
ONLINE-B	*80.0*	89	9	0	2
Gemma-3-12B	82.0	89	7	0	4
Qwen3-235B	82.0	90	8	0	2
UvA-MT	82.0	91	9	0	0
hybrid	83.0	91	8	0	1
Qwen2.5-7B	84.0	91	7	0	2
SalamandraTA	85.0	91	6	0	3
CommandA	86.0	93	7	0	0
DeepSeek-V3	86.0	93	7	0	0
Gemma-3-27B	86.0	93	7	0	0
GPT-4.1	86.0	93	7	0	0
Wenyiil	*86.0*	93	7	0	0
EuroLLM-9B	89.0	93	4	0	3
IRB-MT	88.0	93	5	0	2
Llama-3.1-8B	88.0	93	5	0	2
IR-MultiagentMT	89.0	94	5	1	0
Llama-4-Maverick	89.0	94	5	0	1
TowerPlus-72B	92.0	96	4	0	0
CommandA-MT	94.0	97	3	0	0
Mistral-7B	94.0	97	3	0	0
TranssionMT	94.0	94	0	0	6
ONLINE-G	98.0	98	0	0	2
TranssionTranslate	100.0	100	0	0	0

Table 24: MUSICAL INSTRUMENTS, Serbian

PET SUPPLIES, English→Russian					
		distribution			
system	score	m	f	x	mix
Yandex	*-63.0*	17	80	3	0
Algharb	*-61.0*	19	80	1	0
Gemini-2.5-Pro	*-52.0*	24	76	0	0
Yolu	*-50.0*	24	74	2	0
Shy	*-41.0*	28	69	2	1
Wenyiil	*-39.0*	30	69	1	0
GemTrans	-37.0	31	68	1	0
TowerPlus-9B	-35.0	32	67	0	1
GPT-4.1	-29.0	35	64	1	0
ONLINE-G	*-29.0*	20	49	1	30
Claude-4	-22.0	38	60	2	0
hybrid	-18.0	38	56	2	4
IRB-MT	-18.0	40	58	2	0
DLUT_GTCOM	-15.0	40	55	1	4
DeepSeek-V3	-12.0	43	55	2	0
Laniqo	-9.0	44	53	1	2
SalamandraTA	-9.0	42	51	1	6
AyaExpanse-32B	-6.0	47	53	0	0
UvA-MT	-6.0	46	52	2	0
Gemma-3-12B	-5.0	46	51	2	1
Qwen3-235B	2.0	50	48	1	1
Gemma-3-27B	11.0	55	44	1	0
ONLINE-W	11.0	35	24	0	41
TowerPlus-72B	13.0	56	43	0	1
CommandA	15.0	57	42	1	0
EuroLLM-22B	16.0	56	40	1	3
Qwen2.5-7B	24.0	53	29	6	12
TranssionTranslate	29.0	51	22	0	27
Llama-4-Maverick	33.0	66	33	1	0
IR-MultiagentMT	35.0	65	30	3	2
AyaExpanse-8B	38.0	68	30	1	1
CommandR7B	38.0	66	28	3	3
ONLINE-B	38.0	68	30	0	2
Llama-3.1-8B	43.0	68	25	0	7
CommandA-MT	50.0	74	24	0	2
SRPOL	53.0	73	20	1	6
EuroLLM-9B	60.0	78	18	2	2
NLLB	76.0	86	10	1	3
Mistral-7B	93.0	93	0	0	7
TranssionMT	98.0	98	0	0	2

Table 25: PET SUPPLIES, Russian

PET SUPPLIES, English→Serbian					
		distribution			
system	score	m	f	x	mix
Gemini-2.5-Pro	*-66.0*	17	83	0	0
Algharb	*-61.0*	19	80	0	1
ONLINE-B	*-47.0*	25	72	0	3
Wenyiil	*-47.0*	26	73	0	1
Shy	*-43.0*	28	71	0	1
GemTrans	*-39.0*	29	68	0	3
Yolu	*-39.0*	28	67	0	5
GPT-4.1	-19.0	40	59	0	1
hybrid	-19.0	39	58	1	2
Claude-4	-6.0	46	52	0	2
AyaExpanse-32B	8.0	50	42	0	8
Gemma-3-27B	8.0	52	44	0	4
IRB-MT	11.0	54	43	0	3
EuroLLM-22B	12.0	53	41	0	6
DeepSeek-V3	15.0	57	42	0	1
CUNI-SFT	16.0	54	38	0	8
TowerPlus-9B	17.0	50	33	4	13
UvA-MT	17.0	57	40	0	3
Gemma-3-12B	19.0	54	35	0	11
Llama-3.1-8B	21.0	58	37	0	5
Qwen3-235B	32.0	66	34	0	0
AyaExpanse-8B	33.0	60	27	2	11
CommandR7B	38.0	58	20	2	20
Qwen2.5-7B	40.0	62	22	1	15
IR-MultiagentMT	45.0	72	27	0	1
CommandA	48.0	71	23	0	6
SalamandraTA	48.0	71	23	0	6
Llama-4-Maverick	49.0	73	24	0	3
EuroLLM-9B	56.0	73	17	0	10
TowerPlus-72B	60.0	77	17	2	4
CommandA-MT	85.0	92	7	0	1
ONLINE-G	89.0	89	0	0	11
Mistral-7B	90.0	92	2	1	5
TranssionMT	93.0	94	1	0	5
TranssionTranslate	100.0	100	0	0	0

Table 26: PET SUPPLIES, Serbian

SPORTS AND OUTDOORS, English→Russian					
		distribution			
system	score	m	f	x	mix
SalamandraTA	5.0	49	44	1	6
Algharb	*8.0*	53	45	1	1
Yolu	*12.0*	55	43	1	1
Gemini-2.5-Pro	*17.0*	58	41	1	0
Wenyiil	*17.0*	57	40	1	2
Laniqo	20.0	59	39	2	0
Shy	*22.0*	59	37	3	1
Yandex	*22.0*	59	37	2	2
ONLINE-G	*24.0*	51	27	0	22
TowerPlus-9B	33.0	66	33	1	0
Claude-4	34.0	65	31	2	2
GemTrans	35.0	67	32	0	1
hybrid	35.0	65	30	4	1
GPT-4.1	37.0	68	31	1	0
EuroLLM-22B	41.0	70	29	0	1
IRB-MT	41.0	70	29	1	0
SRPOL	44.0	71	27	1	1
AyaExpanse-32B	45.0	72	27	0	1
CommandA	46.0	73	27	0	0
DeepSeek-V3	46.0	72	26	2	0
Gemma-3-12B	46.0	72	26	0	2
DLUT_GTCOM	47.0	71	24	3	2
UvA-MT	47.0	73	26	0	1
Qwen3-235B	49.0	74	25	0	1
IR-MultiagentMT	53.0	76	23	1	0
TowerPlus-72B	54.0	76	22	0	2
Gemma-3-27B	56.0	77	21	1	1
ONLINE-W	56.0	66	10	0	24
CommandR7B	58.0	76	18	3	3
AyaExpanse-8B	60.0	79	19	0	2
Llama-3.1-8B	60.0	79	19	0	2
Qwen2.5-7B	60.0	75	15	9	1
ONLINE-B	62.0	78	16	1	5
CommandA-MT	65.0	82	17	0	1
Llama-4-Maverick	65.0	82	17	1	0
EuroLLM-9B	66.0	83	17	0	0
TranssionTranslate	69.0	78	9	0	13
NLLB	80.0	88	8	0	4
Mistral-7B	90.0	94	4	0	2
TranssionMT	98.0	98	0	0	2

Table 27: SPORTS AND OUTDOORS, Russian

SPORTS AND OUTDOORS, English→Serbian					
		distribution			
system	score	m	f	x	mix
Gemini-2.5-Pro	*4.0*	52	48	0	0
Algharb	*6.0*	53	47	0	0
Wenyiil	*10.0*	55	45	0	0
EuroLLM-22B	16.0	56	40	0	4
Yolu	*22.0*	60	38	0	2
CUNI-SFT	26.0	58	32	0	10
Shy	*27.0*	63	36	0	1
GemTrans	*28.0*	63	35	0	2
ONLINE-B	*31.0*	64	33	0	3
GPT-4.1	34.0	67	33	0	0
Claude-4	36.0	68	32	0	0
AyaExpanse-32B	37.0	65	28	1	6
IRB-MT	37.0	68	31	0	1
hybrid	38.0	69	31	0	0
TowerPlus-9B	40.0	65	25	1	9
DeepSeek-V3	41.0	70	29	0	1
Gemma-3-12B	41.0	69	28	0	3
AyaExpanse-8B	42.0	65	23	1	11
CommandR7B	45.0	63	18	4	15
UvA-MT	45.0	72	27	0	1
Gemma-3-27B	48.0	72	24	0	4
IR-MultiagentMT	48.0	73	25	2	0
Llama-3.1-8B	53.0	73	20	0	7
Qwen3-235B	57.0	78	21	0	1
EuroLLM-9B	58.0	74	16	1	9
CommandA	60.0	79	19	0	2
Qwen2.5-7B	64.0	75	11	3	11
SalamandraTA	68.0	83	15	0	2
Llama-4-Maverick	69.0	84	15	0	1
CommandA-MT	70.0	84	14	0	2
TowerPlus-72B	71.0	84	13	0	3
Mistral-7B	81.0	85	4	0	11
ONLINE-G	89.0	91	2	0	7
TranssionMT	92.0	94	2	0	4
TranssionTranslate	100.0	100	0	0	0

Table 28: SPORTS AND OUTDOORS, Serbian

Tools, English→Russian

		distribution			
system	score	m	f	x	mix
Yolu	*29.0*	63	34	1	2
Gemini-2.5-Pro	*33.0*	66	33	1	0
Yandex	*36.0*	68	32	0	0
Algharb	*38.0*	68	30	2	0
ONLINE-G	*39.0*	58	19	1	22
Laniqo	41.0	67	26	5	2
TowerPlus-9B	41.0	70	29	0	1
SalamandraTA	47.0	71	24	3	2
Wenyiil	*48.0*	73	25	2	0
Shy	*50.0*	74	24	2	0
GPT-4.1	54.0	76	22	2	0
hybrid	56.0	75	19	5	1
GemTrans	58.0	79	21	0	0
DeepSeek-V3	59.0	79	20	1	0
IRB-MT	59.0	79	20	1	0
AyaExpanse-32B	61.0	78	17	4	1
Gemma-3-12B	61.0	80	19	1	0
Claude-4	62.0	80	18	2	0
Qwen3-235B	62.0	80	18	2	0
EuroLLM-22B	64.0	80	16	3	1
TowerPlus-72B	66.0	82	16	1	1
UvA-MT	66.0	82	16	2	0
ONLINE-W	67.0	76	9	1	14
AyaExpanse-8B	68.0	84	16	0	0
CommandA	69.0	83	14	3	0
IR-MultiagentMT	69.0	82	13	5	0
Gemma-3-27B	70.0	85	15	0	0
SRPOL	70.0	84	14	1	1
DLUT_GTCOM	71.0	84	13	2	1
Llama-3.1-8B	72.0	82	10	0	8
Llama-4-Maverick	77.0	87	10	3	0
ONLINE-B	77.0	87	10	1	2
TranssionTranslate	77.0	84	7	1	8
CommandR7B	79.0	87	8	2	3
Qwen2.5-7B	80.0	87	7	2	4
CommandA-MT	83.0	91	8	1	0
EuroLLM-9B	85.0	91	6	3	0
NLLB	91.0	94	3	2	1
Mistral-7B	95.0	96	1	0	3
TranssionMT	99.0	99	0	0	1

Table 29: Tools and Home Improvement, Russian

Tools, English→Serbian

		distribution			
system	score	m	f	x	mix
Gemini-2.5-Pro	*32.0*	66	34	0	0
ONLINE-B	*33.0*	66	33	0	1
Algharb	*34.0*	67	33	0	0
CUNI-SFT	38.0	65	27	1	7
GemTrans	*38.0*	69	31	0	0
Yolu	*41.0*	70	29	0	1
EuroLLM-22B	43.0	69	26	0	5
Wenyiil	*44.0*	72	28	0	0
Shy	*46.0*	73	27	0	0
AyaExpanse-32B	47.0	70	23	0	7
GPT-4.1	51.0	75	24	0	1
CommandR7B	52.0	70	18	3	9
IRB-MT	52.0	76	24	0	0
TowerPlus-9B	52.0	72	20	2	6
Gemma-3-12B	53.0	75	22	0	3
Gemma-3-27B	53.0	76	23	1	0
hybrid	54.0	77	23	0	0
Claude-4	56.0	78	22	0	0
AyaExpanse-8B	57.0	73	16	0	11
UvA-MT	62.0	81	19	0	0
DeepSeek-V3	66.0	83	17	0	0
CommandA	71.0	84	13	0	3
IR-MultiagentMT	72.0	85	13	0	2
Llama-3.1-8B	72.0	85	13	0	2
Qwen3-235B	73.0	86	13	0	1
Qwen2.5-7B	74.0	81	7	0	12
EuroLLM-9B	75.0	81	6	0	13
Llama-4-Maverick	77.0	88	11	0	1
TowerPlus-72B	79.0	88	9	0	3
CommandA-MT	80.0	90	10	0	0
SalamandraTA	80.0	88	8	0	4
Mistral-7B	90.0	93	3	0	4
ONLINE-G	91.0	91	0	1	8
TranssionMT	92.0	92	0	0	8
TranssionTranslate	100.0	100	0	0	0

Table 30: Tools and Home Improvement, Serbian

VIDEO GAMES, English→Russian

		distribution			
system	score	m	f	x	mix
Yandex	*60.0*	79	19	2	0
Yolu	*63.0*	80	17	1	2
SalamandraTA	65.0	81	16	1	2
Laniqo	69.0	83	14	2	1
Shy	*75.0*	85	10	5	0
Gemini-2.5-Pro	*77.0*	87	10	3	0
TowerPlus-9B	77.0	88	11	1	0
Gemma-3-12B	78.0	87	9	3	1
Algharb	*79.0*	89	10	1	0
ONLINE-G	*79.0*	85	6	1	8
hybrid	80.0	88	8	4	0
IRB-MT	80.0	89	9	2	0
TowerPlus-72B	81.0	90	9	1	0
UvA-MT	82.0	91	9	0	0
Wenyiil	*83.0*	90	7	3	0
Claude-4	84.0	91	7	2	0
Qwen2.5-7B	84.0	88	4	7	1
Qwen3-235B	84.0	92	8	0	0
IR-MultiagentMT	85.0	91	6	3	0
DeepSeek-V3	86.0	92	6	2	0
ONLINE-W	87.0	90	3	3	4
EuroLLM-22B	88.0	93	5	1	1
Gemma-3-27B	88.0	93	5	2	0
GPT-4.1	88.0	93	5	2	0
SRPOL	89.0	94	5	0	1
AyaExpanse-32B	90.0	95	5	0	0
AyaExpanse-8B	90.0	95	5	0	0
GemTrans	90.0	95	5	0	0
NLLB	90.0	92	2	4	2
CommandR7B	92.0	95	3	1	1
CommandA	93.0	96	3	1	0
TranssionTranslate	93.0	94	1	2	3
CommandA-MT	94.0	96	2	2	0
Llama-4-Maverick	94.0	97	3	0	0
DLUT_GTCOM	95.0	97	2	1	0
EuroLLM-9B	97.0	98	1	1	0
ONLINE-B	98.0	98	0	2	0
Mistral-7B	99.0	99	0	0	1
TranssionMT	99.0	99	0	1	0
Llama-3.1-8B	100.0	100	0	0	0

Table 31: VIDEO GAMES, Russian

VIDEO GAMES, English→Serbian

		distribution			
system	score	m	f	x	mix
CUNI-SFT	52.0	74	22	1	3
AyaExpanse-8B	57.0	74	17	0	9
AyaExpanse-32B	67.0	80	13	1	6
CommandR7B	71.0	80	9	6	5
EuroLLM-22B	71.0	83	12	0	5
TowerPlus-9B	73.0	83	10	1	6
Yolu	*77.0*	87	10	0	3
Algharb	*78.0*	89	11	0	0
Shy	*78.0*	89	11	0	0
Wenyiil	*78.0*	89	11	0	0
Claude-4	80.0	90	10	0	0
Gemini-2.5-Pro	*80.0*	90	10	0	0
Gemma-3-12B	80.0	88	8	0	4
CommandA	81.0	89	8	0	3
GemTrans	*81.0*	90	9	0	1
IRB-MT	81.0	90	9	0	1
DeepSeek-V3	84.0	92	8	0	0
EuroLLM-9B	87.0	92	5	0	3
Qwen2.5-7B	87.0	88	1	0	11
UvA-MT	87.0	93	6	0	1
GPT-4.1	88.0	94	6	0	0
ONLINE-B	*88.0*	94	6	0	0
Gemma-3-27B	89.0	92	3	0	5
Llama-3.1-8B	90.0	94	4	0	2
SalamandraTA	90.0	91	1	1	7
TowerPlus-72B	91.0	95	4	1	0
hybrid	92.0	96	4	0	0
Qwen3-235B	92.0	96	4	0	0
IR-MultiagentMT	94.0	96	2	0	2
TranssionMT	94.0	95	1	1	3
ONLINE-G	96.0	96	0	0	4
Mistral-7B	97.0	98	1	0	1
CommandA-MT	98.0	99	1	0	0
Llama-4-Maverick	98.0	99	1	0	0
TranssionTranslate	100.0	100	0	0	0

Table 32: VIDEO GAMES, Serbian