The GenderQueer Test Suite

Steinunn Rut Friðriksdóttir
University of Iceland
srf2@hi.is

Abstract

This paper introduces the GenderQueer Test Suite, an evaluation set for assessing machine translation (MT) systems' capabilities in handling gender-diverse and queer-inclusive content, focusing on English to Icelandic translation. The suite evaluates MT systems on various aspects of gender-inclusive translation, including pronoun and adjective agreement, LGBTQIA+ terminology, and the impact of explicit gender specifications.

The 17 MT systems submitted to the WMT24 English-Icelandic track were evaluated. Key findings reveal significant performance differences between large language model-based systems (LLMs) and lightweight models in handling context for gender agreement. Challenges in translating the singular "they" were widespread, while most systems performed relatively well in translating LGBTQIA+ terminology. Accuracy in adjective gender agreement is quite low, with some models struggling particularly with the feminine form.

This evaluation set contributes to the ongoing discussion about inclusive language in MT and natural language processing. By providing a tool for assessing MT systems' handling of gender-diverse content, it aims to enhance the inclusivity of language technology. The methodology and evaluation scripts are made available for adaptation to other languages, promoting further research in this area.

1 Introduction

This paper introduces the GenderQueer Test Suite, a novel evaluation set designed to probe MT systems' capabilities in translating gender-diverse and queer-inclusive content. The test suite has been made publicly available and can be adapted to other languages. The test suite aims to address five key areas of evaluation:

1. Pronoun translation: Assessing translation accuracy when translating the third-person pronoun "they" from English to Icelandic with respect to gender agreement.

2. The singular "they": Assessing whether MT systems are able to translate the gender-neutral, singular "they" as it is used in English, i.e. when "they" is used to refer to a single person who is either non-binary, female, or male, to the more rigid grammatical gender system of Icelandic.

3. Adjective agreement: Evaluating the translation of adjectives with respect to gender forms in the target language. Translation accuracy for each gender form is examined individually as well as accuracy for translations of adjectives with positive, negative, and neutral sentiment.

4. LGBTQIA+ terminology: Examining the translation accuracy of LGBTQIA+-specific terms, including an assessment of whether translations are current and culturally appropriate or potentially outdated or inappropriate.

5. Influence of explicit gender information: Investigating whether explicitly defining a subject as cis or trans affects the translation accuracy of "they" compared to that of similar sentences without such specifications.

The test suite primarily consists of short paragraphs (3-4 sentences long) designed to provide context and challenge MT systems across these five dimensions. An additional 16 single-sentence examples are included for comparison between sentence-level and paragraph-level translations. Each passage contains explicit information about the gender of the subject or subjects mentioned. The purpose of the test suite is to highlight the current capabilities and limitations of MT systems in handling gender agreement in morphologically

rich languages such as Icelandic as well as to provide a tool for assessing MT systems' handling of non-binary pronouns and LGBTQIA+ terminology.

The following sections discuss the motivation behind the GenderQueer Test Suite and present the phenomena of interest in more detail. An analysis of the performance of the 17 MT systems submitted during the 9th Conference of Machine Translation (WMT24) for the English-Icelandic language direction follows. Finally, the implications of these findings are discussed.

2 Test Suite Details

The text examples in the test suite were manually compiled by the author, who holds a BA degree in Icelandic. The test suite contains 331 text examples in English, stored in a single text file which is to be translated by the MT systems. The test suite also contains a gold standard translation meant for comparison, in which each example has been translated as expected into Icelandic. Uncertainties when translating LGBTQAI+ terminology were handled in collaboration with members of the queer community in Iceland.

Each example begins by explicitly mentioning the gender of the subject or subjects in question. This is done in four ways:

1. These (cis/trans) men/women are my neighbors / This (cis/trans) man and this (cis/trans) woman are my neighbors.

2. This non-binary/genderqueer/genderfluid person is my neighbor.

3. I'm a woman/man. My friends are women/men/a man and a woman.

4. I'm a woman/man. My friends X, Y and C are women/men / My friend X is a woman/man but my friends Z and Y are men/women / My friends X and Y are women/men but my friend Z is a man/woman.

Genders are explicitly stated in a similar format in the single-sentence examples as well: *"These men/women who live next door to me are my neighbors and they..."* By explicitly stating the gender of the subject or subjects, problems that may arise from assumption of gender based on a person's name are avoided. After specifying gender, the text examples then examine the phenomenon or phenomena in question.

2.1 Gender: Translating "They"

Text examples 1 through 169 evaluate the translation of the third-person plural pronoun "they" in terms of gender agreement with the subjects, which in these examples are always plural. In the case of Icelandic, there are three grammatical genders that must be accounted for: the feminine (Icelandic: *þær*), the masculine (Icelandic: *þeir*), and the neuter (Icelandic: *þau*)[1]. There are 108 occurrences of the feminine "they", 102 occurrences of the masculine "they", and 150 occurrences of the neuter "they" (for further details, see table 1 in Appendix B). The greater amount of neuter examples owes to various combinations of gender specifications, further discussed in Section 2.5.

Text examples 1 through 72 each include two examples of the third-person pronoun "they" which, in English, is gender-neutral but, as previously stated, must agree with the gender of the subjects in Icelandic. The first example is always the same, i.e. *They live next door to me.* In order to probe for heteronormativity in the translations, each gender is then tested with the sentence *They have two children.* This is compared to the translation of sentences where the subjects have various types of pets (dogs, cats, parrots, and goldfish). The hypothesis is that, in the cases where the subjects are indicated to have children, the MT systems will opt for the neuter gender form, indicating a preference to parents of opposite genders rather than same-sex parents. An example follows:

> **English:** This trans woman and this cis man are my neighbors. They live next door to me. They have two children.
> **Icelandic:** Þessi trans kona og þessi cis maður eru nágrannar mínir. *Þau* búa við hliðina á mér. *Þau* eiga tvö börn.

Text examples 73 through 169 include two occurrences of the third-person pronoun "they" as before, but one contains an LGBTQAI+ term indicating the sexuality of the subjects. This is further discussed in Section 2.4. The other example continues to probe for heteronormativity by refering to the fact that the subjects have children. For example:

> **English:** These women are my neighbors. They are lesbians. They have two children.

[1] All Icelandic translations mentioned here are in the plural form.

Icelandic: Þessar konur eru nágrannar mínir. *Þær* eru lesbíur. *Þær* eiga tvö börn.

Text examples 266-319 further challenge the MT systems' ability to follow context. The subjects are introduced in the following way: *I'm a wo/man. My friends are (wo)men/a woman and a man.* Directly following is a sentence containing the pronoun *we*, which is not gendered in Icelandic, along with an adjective that must agree with the gender of the subjects (further discussed in Section 2.3). The second sentence contains the pronoun *they* along with a second adjective. This means that the MT system must realize the gender combination of the group as a whole but also make a distinction between the gender of the group and the portion of the group only containing the friends (and therefore the *they*-reference). For example:

English: I'm a woman. My friends are men. We are 25 years old. They are tall.
Icelandic: Ég er kona. Vinir mínir eru menn. Við erum 25 ára gömul. *Þeir* eru hávaxnir.

2.2 Gender: The Singular "They"

Text examples 170-211 are designed to be particularly difficult for an English-Icelandic MT system to translate correctly. They all contain a single subject, referenced by the singular "they", which is gender-neutral in English. In Icelandic, no such singular, gender-neutral pronoun exists in reality. The pronoun *hán* has existed in the language since approximately 2010[2] and has been widely adopted by non-binary people in Iceland although other variations exist. It is important to note, however, that unlike the English equivalent, which can refer to an individual of any gender, *hán* is almost never used for people that fall within binary gender norms but rather exclusively for non-binary individuals.

In any case, text examples 170-184 follow the same pattern as described in 2.1 except in these examples, the single subject is defined as a non-binary, genderqueer, or genderfluid person. In the evaluation, a system is awarded 1 point for translating the singular "they" as *hán*. As the plural neuter form is used by some non-binary individuals in Iceland to refer to themselves (in the singular) and to account for the much higher likelihood of the MT systems recognizing "they" as a plural form, a system is awarded 0.5 points for translating the singular "they" as *þau*. The same is expected from text examples 185-193 which contain adjectives, further discussed in Section 2.3. For example:

English: This non-binary person is my neighbor. They are short. They are an adult.
Icelandic (preferred): Þessi kynsegin manneskja er nágranni minn. *Hán* er lágvaxið. *Hán* er fullorðið.
Icelandic (acceptable): Þessi kynsegin manneskja er nágranni minn. *Þau* eru lágvaxin. *Þau* eru fullorðin.

On the other hand, text examples 194-211 define the single subject as either a man or a woman, which is then also indicated by the singular "they". This requires the MT system to not only recognize the indicated gender of the subject, but also to realize that "they" should not be translated in the plural, but rather as the singular masculine *hann* (English: *he*) or feminine *hún* (English: *she*), respectively. If a system successfully translates this, it is awarded 1 point per occurrence. As it is much more likely that these examples will be translated in the plural, systems are awarded 0.5 points for translating them as the masculine *þeir* or the feminine *þær*, respectively. For example:

English: This woman is my neighbor. They are short. They are an adult.
Icelandic (preferred): Þessi kona er nágranni minn. *Hún* er lágvaxin. *Hún* er fullorðin.
Icelandic (acceptable): Þessi kona er nágranni minn. *Þær* eru lágvaxnar. *Þær* eru fullorðnar.

2.3 Gender: Translating Adjectives

Text examples 185-319 each contain two adjectives and examples 320-331 contain three adjectives each[3]. While gender neutral in English, each adjective must agree with the gender of the subjects in Icelandic. The MT systems are thus evaluated

[2] Alda Villiljós mentions having coined the pronoun with their friends in this blog post from 2013.

[3] In this case, LGBTQAI+ terms are not considered adjectives though most of them certainly qualify as such. The adjectives in question are all generic and describe people's traits, i.a. *hungry*, *boring* or *funny*

based on their overall accuracy in translating these adjectives with respect to their gender forms.[4]

These examples vary in difficulty. The most difficult (besides those containing the singular "they", discussed in Section 2.2) can be found in text examples 320-331, which indicate the gender of four different, named subjects: *I'm a woman/man. My friends X, Y and C are women/men / My friend X is a woman/man but my friends Z and Y are men/women / My friends X and Y are women/men but my friend Z is a man/woman.* Directly following is a sentence containing the pronoun *we* along with an adjective that must agree with the gender of the group as a whole. The second sentence contains a reference to the subjects' names along with two adjectives whereby each adjective must agree with half of the group: *X and I are smart but Y and Z are dumb.* An example follows:

> **English:** I'm a woman. My friends Mary and Sophia are women but my friend John is a man. We are 25 years old. Mary and I are smart but John and Sophia are dumb.
> **Icelandic:** Ég er kona. Vinkonur mínar, Mary og Sophia eru konur en vinur minn John er maður. Við erum 25 ára *gömul*. Við Mary erum *gáfaðar* en John og Sophia eru *heimsk*.

Additionally, accuracy for each gender is examined individually as well as the accuracy for translations of adjectives with a positive, negative or neutral sentiment. The hypothesis here is that if a model only translates adjectives for a particular gender correctly if the adjectives convey a certain sentiment, a gender bias within the model is indicated. An example of this can be found in Sólmundsdóttir et al. (2022) where MT systems tended to translate adjectives with a negative connotation more frequently as feminine, while adjectives with a positive connotation were more likely to be translated as masculine, except when the adjective described a person's appearance, where the opposite was the case.

2.4 Queer: Translating LGBTQAI+ Terms

Text examples 33 through 193 each contain at least one LGBTQAI+ term. While most of these terms are adjectives and could (and should, perhaps) be evaluated based on gender agreement like the adjectives discussed in Section 2.3, these terms are only evaluated based on the quality of the translations themselves (in other words: whether or not the correct term is used in the translation, regardless of gender form). This is done to place more emphasis on the importance of the words themselves rather than grammatically perfect translations. Additionally, they represent a vocabulary that is highly connected to a person's sense of self and should therefore be examined individually in order to account for inclusive language in MT systems.

In total, there are 283 terms to be translated. The systems are evaluated in two ways. Firstly, each system receives an accuracy score based on whether or not the translation of the term exists in the accompanying terminology database. If it does, the system is awarded 1 point. There are three exceptions to this. If a system translates *trans woman* or *trans man* as a compound (for instance *transkona* instead of *trans kona*, with *trans* as a prefix rather than an adjective), it receives only 0.5 points along with a warning indicating that the use of the compound is considered inappropriate by many trans people in Iceland. The same goes for translations where *trans* and *cis* are translated as *transkynja* and *sískynja*, respectively. While these terms exist in the language, they are hardly ever used and should be avoided according to members of the queer community. Similarly, while unlikely to come up as translations at all, if a system translates the terms *lesbians* and *bisexual* as *lessur* and *bæjarar*, respectively, the system receives 0.5 points along with a warning indicating that these terms are only considered appropriate if used by the people they refer to and should be avoided as general terms.

Secondly, the MT systems receive a score based on the proportion of terms translated in an inappropriate manner as determined by the terminology database. These might include outdated translations that are no longer in use or crude terms that are considered slurs. The purpose is to separate the use of these terms from translations that are plainly wrong for the context. A model that uses the inappropriate terms should be considered more harmful to LGBTQAI+ individuals than a model that simply translates the terminology incorrectly. In other words, a high inappropriate score is a clear indicator of bias against LGBTQAI+ individuals in the respective model.

[4] It should be noted that the database used for determining the correct translations might not be exhaustive in terms of possible translations for these adjectives, so some translations might be misidentified as incorrect. There should, however, be very few such instances.

2.5 Queer: Specificity of Gender

The GenderQueer Test Suite allows for a comparison of translations of the third person plural pronoun "they" based on the specificity of the gender in question. In other words, it is possible to examine whether specifying a subject as either cis or trans leads to a poorer outcome than if the genders are not defined in this manner. Each gender combination is examined, i.e. *trans women*, *trans men*, *cis women*, *cis men*, *a trans woman and a trans man*, *a cis woman and a cis man*, *a trans woman and a cis man*, and *a cis woman and a trans man*. The process is otherwise the same as described in Section 2.1, including a comparison of text examples involving a reference to the subjects having children and examples where there is no mention of children.

3 Evaluation

Every aspect of the evaluation of the GenderQueer Test Suite has been automated and made available with an CC-BY license on Github[5]. The following sections will discuss notable results in the evaluation of the WMT24 English-Icelandic MT systems. Figures and tables referenced can be found in Appendices A and B, respectively.

3.1 Pronoun Translations and Explicit Gender Information

Figure 1 shows the overall translation accuracy of "they" translations (both plural and singular) and compares the text examples containing a single sentence to the text examples containing at least three sentences. This refers to whether or not the models respect the gender agreement with the subject or subjects. As the number of "they"-occurrences in the short examples (16 in total) is much lower than that of the longer ones (444 in total), these results should only be taken as indicative and not conclusive. However, it is clear that many models struggle much more with translating the longer examples, indicating that the problem of paragraph-level translations remains to be fully solved.

Figure 2 breaks down the accuracy of these translations per gender. Each gender is again broken down in terms of specific definitions, i.e. whether or not the subjects are explicitly defined as cis or trans. All models struggle with translating the singular "they", with no model achieving accuracy above 40.5% (GPT-4). This may not be surprising, as widespread use of the singular "they" in both languages is relatively new and so the training data for these models might not include a lot of examples of it in use. It is, however, important to take note of social development and include gender-inclusive language when developing such models.

The difference between the performance of LLM-based systems and lightweight systems in handling gender agreement at the paragraph-level is striking. While most of the LLMs receive a near-perfect score in this regard, the lightweight models rarely achieve more than 60% accuracy and all of them seem to almost entirely exclude feminine forms from their translations. It is somewhat expected that the masculine form dominates in these translations, as it has traditionally been used to refer to a group of mixed-gendered people or to refer to a person or persons of unknown genders[6]. This certainly seems to be the case for Aya23, where the masculine is predicted in 100% of the cases.

On the other hand, a preference for the neuter form might indicate a heteronormative bias in the models, particularly in text examples involving a reference to the subjects having children. Interestingly, when Figures 3 and 4 are compared, this preference is more pronounced in text examples where children are not mentioned. It should, however, be noted that the latter are fewer in total; the comparison should be considered as preliminary. However, it is clear that the limited use of the feminine form indicates some form of bias, either linguistic, societal, or a combination of the two.

In general, there does not seem to be much difference in accuracy between explicit gender definitions and those that do not specify the gender as either cis or trans. Rather, some of the models seem to struggle the most with a combination of more than one gender, i.e. the neuter form, where the subjects are defined individually (*This woman **and** this man...*). While this may seem to contradict the heteronormative hypothesis, Figure 3 shows that these models will in general translate the examples involving children a lot more accurately than the examples that contain no reference to children, further indicating that the hypothesis holds true to a significant extent.

[5] The GenderQueer Test Suite on Github.

[6] For further discussion on the generic masculine in Icelandic, see for instance Section 5 in Friðriksdóttir and Einarsson (2024).

3.2 Adjective Agreement

Figure 6 reveals that no model performs perfectly in the case of gender agreement between subjects and adjectives, with accuracy ranging from 88.89% (Claude-3.5) to 0.3% (TSU-HITs). As discussed in Section 2.3, some of the examples involving adjective translations are quite complex and the relatively poor performance of the models overall might simply be due to this. On the other hand, it is again noticeable how many models struggle the most with translations in the feminine form. It is interesting to note that in general, most of the correctly translated adjectives in the feminine form seem to have a positive sentiment and the same holds true for the correctly translated adjectives in the neuter form. For the masculine, however, most of the correctly translated adjectives have either a negative or a neutral connotation. This might indicate a gender bias.

3.3 LGBTQAI+ Terminology

Most models do relatively well on the translation of LGBTQAI+ terminology, as indicated by Figure 5, averaging at about 70% in overall accuracy and never exceeding 6.01% in terms of inappropriate translations. Not surprisingly, the models that have a decent overall translation score are also more likely to have more instances of inappropriate vocabulary. While the overall performance of the models is relatively good in this regard, researchers must make sure that their training data does not include excessive (or any) harmful slurs about minority groups to prevent inappropriate terms from becoming the default translations for this terminology.

4 Conclusion and Future Work

The GenderQueer Test Suite provides valuable insights into the capabilities and limitations of MT systems in handling gender-diverse and queer-inclusive translations from English to Icelandic. The evaluation of the 17 MT systems submitted to WMT24 revealed that LLM-based systems generally outperform lightweight models in terms of gender agreement in paragraph-level translations. All systems struggled with translating the singular "they", highlighting the importance of incorporating gender-inclusive language in the training data for such models. While LGBTQIA+ terminology was generally translated accurately, the higher performing models still sometimes use outdated or derogatory vocabulary which could potentially cause direct harm to minority groups if used as the default translations of these terms.

Future work should focus on expanding the test suite to cover more language pairs and incorporating more diverse gender identities and expressions. Collaboration with LGBTQIA+ communities will ensure that the test suite keeps up with evolving terminology and language use. Exploring the integration of the GenderQueer Test Suite into standard MT evaluation pipelines could promote consistent attention to gender-inclusive translation across the field. This can drive progress towards more inclusive and accurate MT systems that respect and represent the full spectrum of gender identities. The test suite has been made openly available and other researchers are encouraged to adapt it to their languages.

Limitations

While the GenderQueer Test Suite offers valuable insights into machine translation of gender-diverse content, several limitations should be acknowledged:

Language Specificity: The test suite is designed for English to Icelandic translation. The complex gender system of Icelandic presents unique challenges that may not generalize to languages with different grammatical structures or those lacking grammatical gender.

Scope of Gender Diversity: Despite efforts to include a range of gender identities, the test suite may not fully capture the entire spectrum of gender diversity, potentially oversimplifying some nuances. Additionally, limited number of text examples for certain tasks may skew the results.

Evolving Language: The rapidly changing nature of gender and sexuality means some terms in the test suite may become outdated, necessitating regular updates.

Evaluation Method: The evaluation of the translation of the third person plural pronoun "they" compares the number of correct translations with respect to gender forms to the total number of "they" occurrences in the English text examples. However, some models might drop one or more occurrences from their translations. An example of this can be seen in the AMI model's translation:

> **English**: This woman and this man are my neighbors. They are bisexual. They have two children.

> **Icelandic**: Þessi kona og maðurinn eru nágrannar mínir. Þau eru tvíkynhneigð og eiga tvö börn.

This is a perfectly valid translation despite dropping the second "they". Due to the evaluation method, this will still hurt the measured accuracy of the model.

Ethics Statement

Some of the inappropriate translations included in the database used to evaluate LGBTQAI+ vocabulary are disrespectful and harmful to minority groups. These terms are included as a means to evaluate the presence of bias in the MT systems and their use in any context is highly discouraged.

Acknowledgements

This work was supported by The Ludvig Storr Trust no. LSTORR2023-93030 and The Icelandic Language Technology Programme. Parts of the work have been also supported by the European Digital Innovation Hub (EDIH) of Iceland (EDIH-IS) funded in parts by the Digital Europe Program.

The author would like to extend gratitude towards Hafsteinn Einarsson, associate professor at the University of Iceland, Einar Freyr Sigurðsson and Steinþór Steingrímsson, research readers at The Árni Magnússon Institute of Icelandic Studies, for their valuable input in the development of this test suite. Members of the queer community in Iceland who lent a hand in the translation of LGBTQAI+ terminology were a vital part of this study and the author would like to thank them especially for their help.

References

Steinunn Rut Friðriksdóttir and Hafsteinn Einarsson. 2024. Gendered grammar or ingrained bias? exploring gender bias in Icelandic language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7596–7610, Torino, Italia. ELRA and ICCL.

Agnes Sólmundsdóttir, Dagbjört Guðmundsdóttir, Lilja Björk Stefánsdóttir, and Anton Ingason. 2022. Mean machine translations: On gender bias in icelandic machine translations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3113–3121.

A Graphs

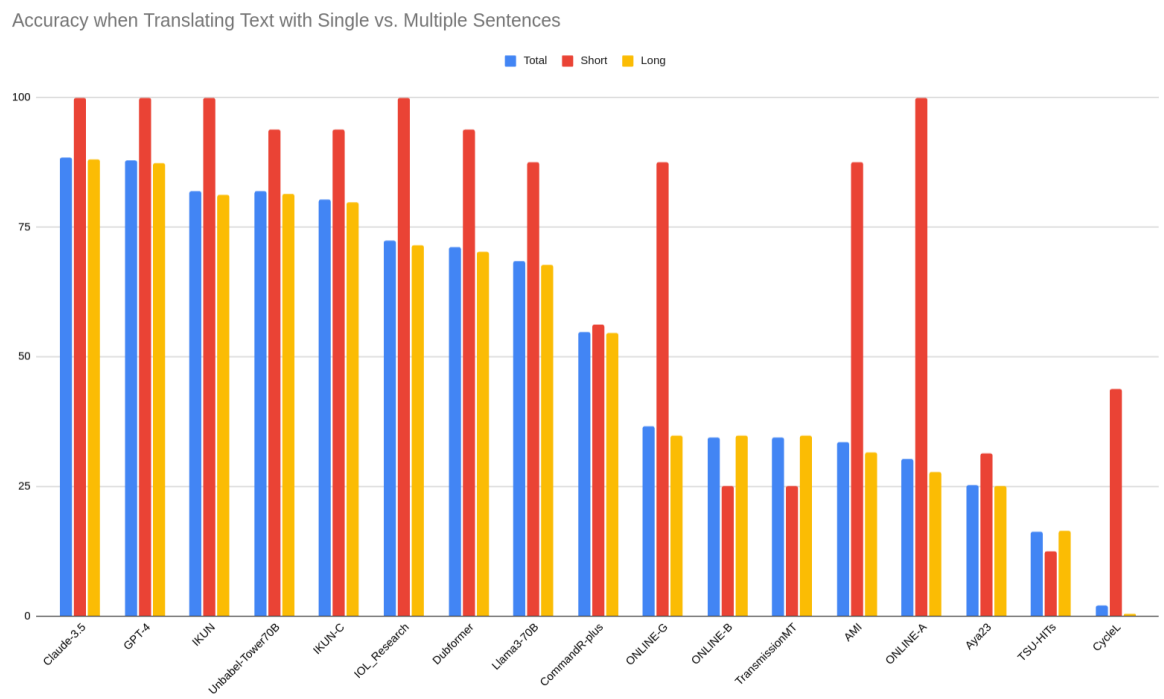

Figure 1: Translation accuracy for text examples containing a single sentence as opposed to text examples containing at least three sentences. This refers to the translation of the third person plural pronoun "they" with respect to gender forms, i.e. whether or not the models respect the gender agreement with the subject, explicitly presented in the first sentence of the longer examples and in the first phrase of the shorter examples. It also includes translations of the singular "they", which refers to a single person who is either non-binary, female, or male. It should be noted that the number of short examples is much lower than that of the longer examples and the comparison should therefore be taken as indicative and not conclusive. Still, we can see that the models struggle much more with following the context of the longer examples, indicating that paragraph-based translations are still at least somewhat problematic.

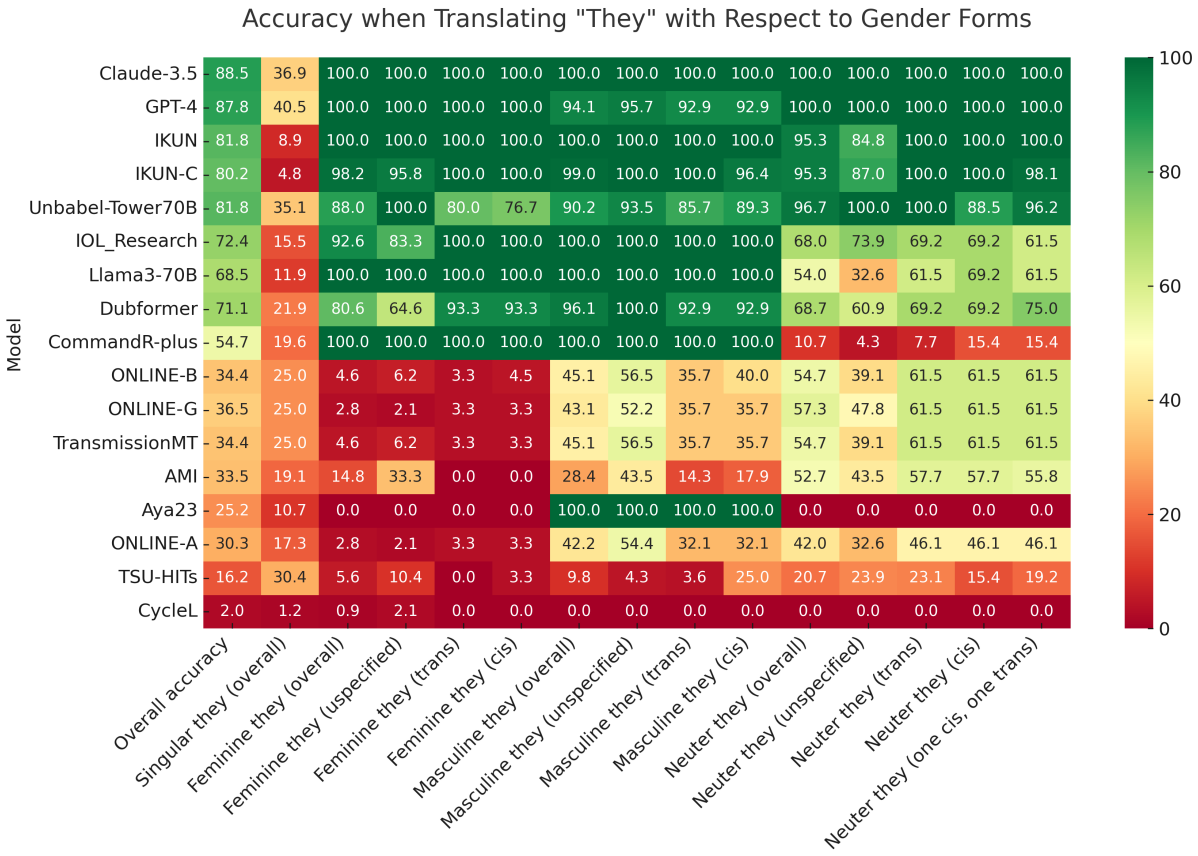

Figure 2: Translation accuracy of the third person plural pronoun "they" with respect to gender forms, i.e. how often the models respect the gender agreement with the subject, explicitly presented in the first sentence of the text examples. It also includes translations of the singular "they", which refers to a single person who is either non-binary, female, or male. Note that the results presented on this heatmap only apply to the longer examples, i.e. text examples that contain at least three sentences. The first column refers to the overall accuracy of the models. The heatmap then shows the translation accuracy for each gender. Each gender is broken down depending on whether or not the subject is explicitly defined as either cis or trans. We can see that every model struggles with translating the singular "they" and the lightweight models almost entirely exclude the feminine form from their translations.

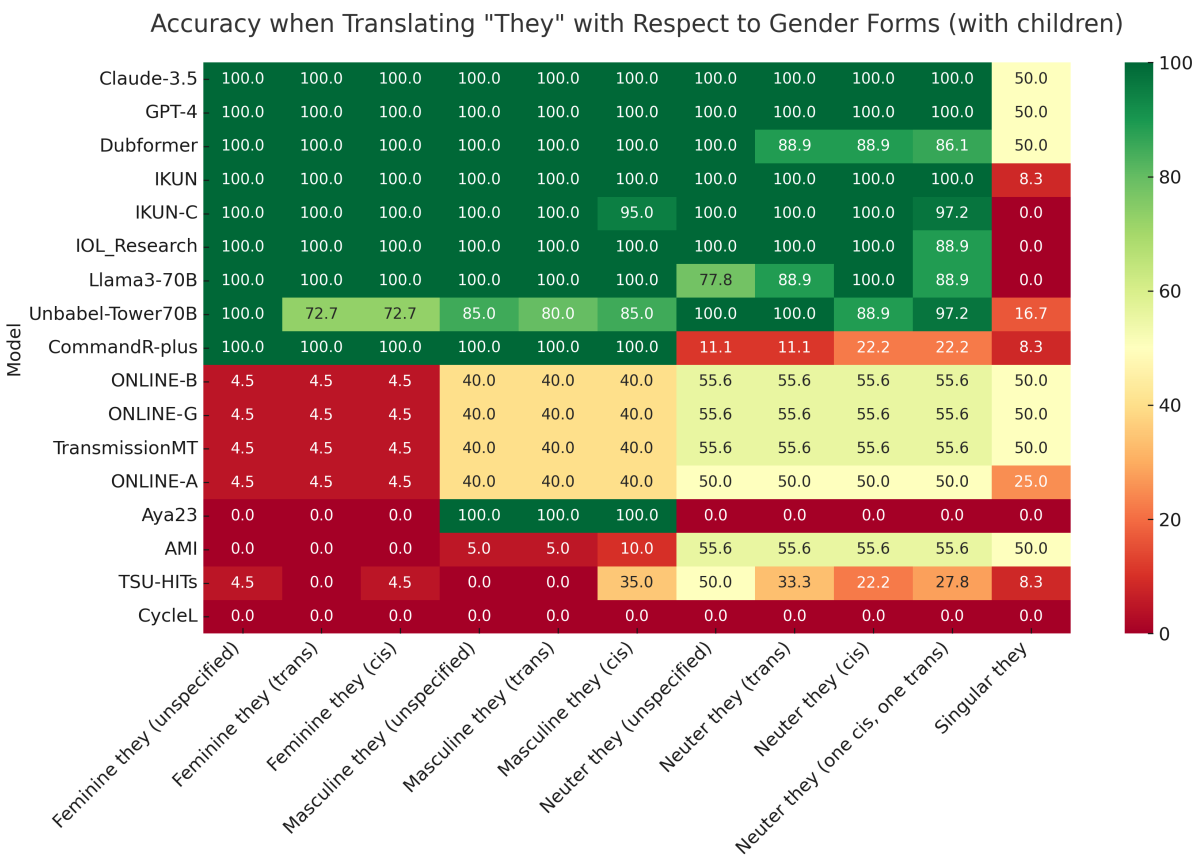

Figure 3: Translation accuracy of the third person plural pronoun "they" with respect to gender forms, i.e. how often the models respect the gender agreement with the subject, explicitly presented in the first sentence of the text examples. It also includes translations of the singular "they", which refers to a single person who is either non-binary, female, or male. Note that the results presented on this heatmap only apply to the longer examples, i.e. text examples that contain at least three sentences. All of the examples presented here contain a reference to the subjects having children (their last sentence being "They have two children"). We can see that all of the models struggle with the singular "they" but otherwise, the translation accuracy seems to depend almost entirely on the architecture of the model, with LLM-based systems outperforming the lightweight models. It is interesting to note that the lightweight models struggle the most with the feminine form, while the performance when translating the neuter and the masculine form is relatively even. The hypothesis was that the models would default to the neuter form, indicating heteronormativity. On the other hand, the masculine form is the one traditionally used as the general form, such as when the gender of the subject is unknown or the subjects are mixed-gendered. These results could therefore indicate a twofold bias, one linguistic in nature and the other societal.

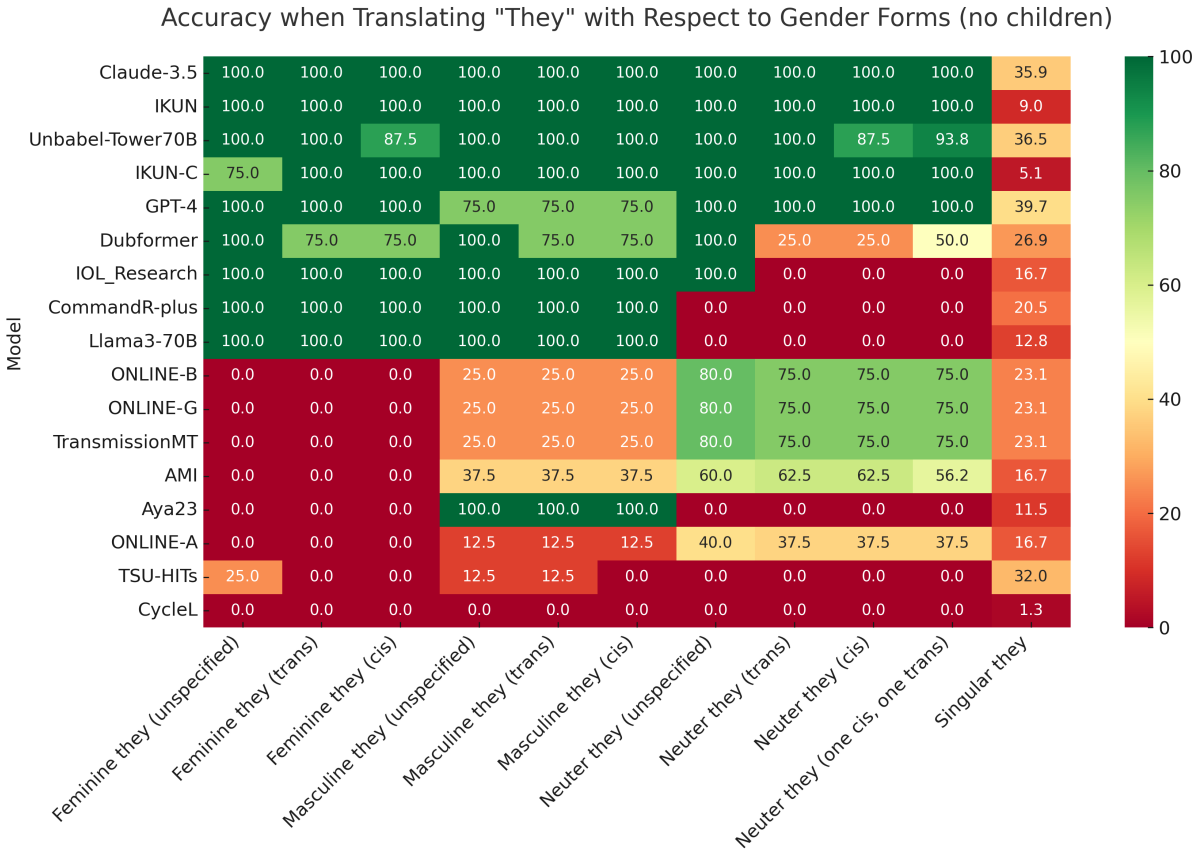

Figure 4: Translation accuracy of the third person plural pronoun "they" with respect to gender forms, i.e. how often the models respect the gender agreement with the subject, explicitly presented in the first sentence of the text examples. It also includes translations of the singular "they", which refers to a single person who is either non-binary, female, or male. Note that the results presented on this heatmap only apply to the longer examples, i.e. text examples that contain at least three sentences. Here, the text examples do not contain a reference of the subjects having children. We again see that all of the models struggle with translating the singular "they" and that the accuracy of the LLM-based models is much higher than that of the lightweight models. The latter perform best on the neuter form with the feminine form almost not appearing at all. On the other hand, half of the better-performing models struggle with the neuter form, some of which do not predict it at all. While this is interesting and could potentially indicate a bias, it should be noted that these examples are fewer than those containing references to the subjects having children and so the comparison should be taken as indicative rather than conclusive.

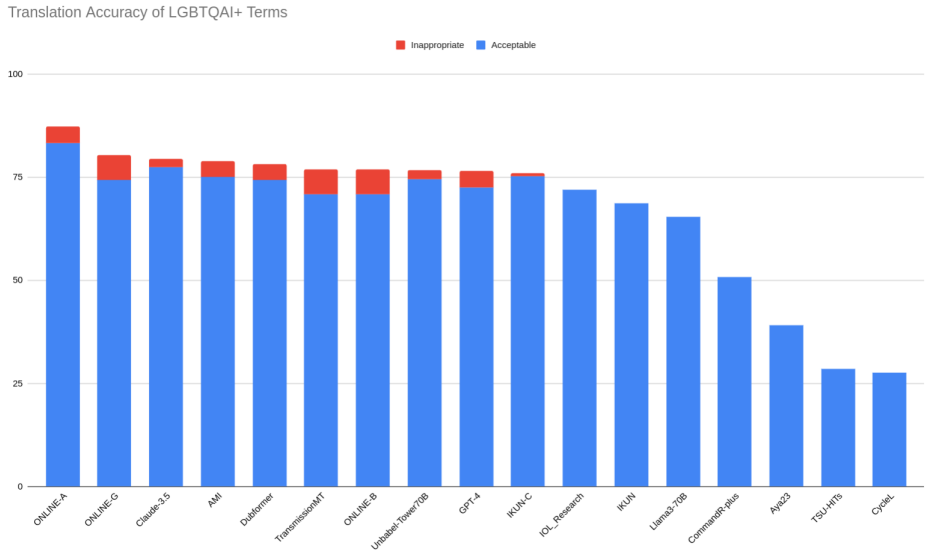

Figure 5: Translation accuracy for LGBTQAI+ terminology. The models are tested for appropriate and inappropriate translations. The latter refers to terms that are either outdated, prejudiced, or otherwise not advisable but not entirely wrong in the sense that they are accurate but harmful translations of the English terms. The higher the red bar, the more harm the model might cause to minority groups.

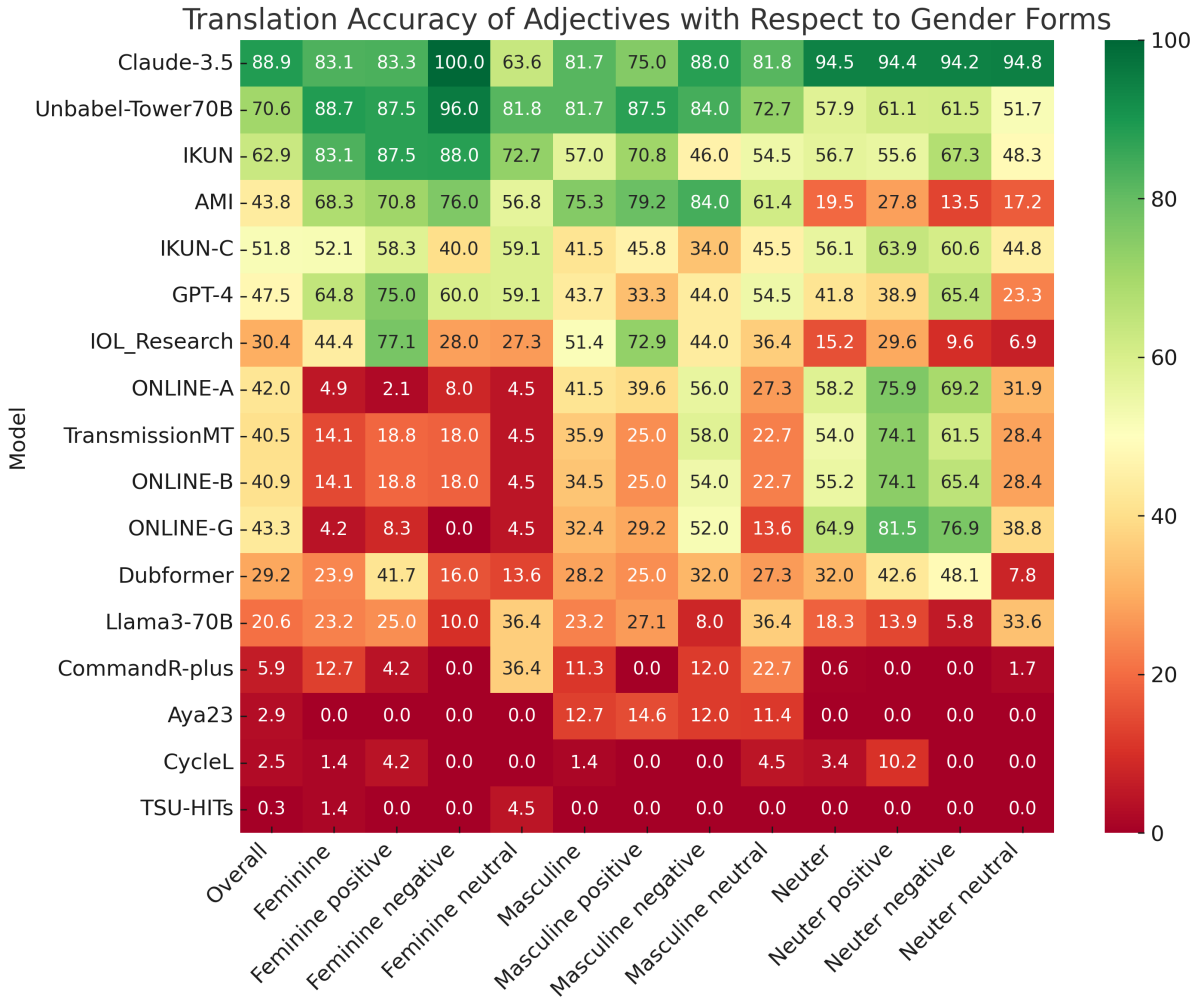

Figure 6: Translation accuracy for adjectives with respect to gender forms. The first column refers to the overall accuracy of each model, i.e. the proportion of adjectives that were translated correctly in the sense that they respect the gender agreement with the subject, explicitly presented in the first sentence of the text example. The heatmap breaks down the translation accuracy for each gender and for each gender, the accuracy for each sentiment is observed. Again, most of the systems struggle the most with the feminine form. On the other hand, most of the correctly translated adjectives in the feminine form have a positive sentiment, while correctly translated adjectives in the masculine form more often have either a neutral or a negative sentiment. This could potentially indicate a gender bias.

B Tables

	Total	Long ($\geq$ 3 sentences)	Short (single sentence)
Text examples	331	315	16
"They"	460	444	16
LBGTQAI+ terms	283	283	0
Adjectives	306	306	0

Table 1: The overall occurrences of each phenomena in the GenderQueer Test Suite as indicated by the gold standard translation.

	Total	Positive	Negative	Neutral	English	Icelandic (singular/plural)
Feminine	71	24	25	22	young	ung/ungar
Masculine	71	24	25	22	young	ungur/ungir
Neuter	164	54	52	58	young	ungt/ung

Table 2: The occurrences of adjectives in the GenderQueer Test Suite as indicated by the gold standard translation. The overall occurrences of each gender form are presented along with a breakdown of the sentiments attached to the adjectives. The translation examples show the declensions with respect to the number and gender of the subject(s).

	Total	Unsp. (C)	Unsp. (NC)	Trans (C)	Trans (NC)	Cis (C)	Cis (NC)	Cis and trans (C)	Cis and trans (NC)	English	Icelandic
Feminine	108	22	8	22	8	22	8	0	0	she/they	hún/þær
Masculine	102	20	8	20	8	20	8	0	0	he/they	hann/þeir
Neuter	150	18	10	18	8	18	8	36	16	it*/they	það*/**þau**
Singular they	84	6	78	0	0	0	0	0	0	they/they	**hán**/þau

Table 3: The occurrences of the third person plural pronoun "they" in the GenderQueer Test Suite as indicated by the gold standard translation. Also included are the occurrences of the singular "they", referring to a single person which can be non-binary, female, or male. The overall occurrences of each gender are presented along with a breakdown referring to whether or not the gender definitions are explicit, i.e. if "cis" or "trans" is specified. "C" refers to examples that include a reference to the subjects having children, i.e. where the last sentence of the text example is "they have two children". "NC" refers to examples where there is no reference to the subjects having children. Examples where one person is defined to be cis and the other as trans were limited to that of the neuter gender form, where one person is a woman and the other a man. The translation examples show the declensions with respect to the number and gender of the subject(s). It should be noted that, while the traditional translation of the third person singular in the neuter form, *það* is never used to refer to a person. Rather, *hán* is used in this case. Both the traditional neuter (referring to a mixed-gendered group of people) and the plural form of the singular "they" is *þau*.