FRI Natural language processing course 2021
Project report

University *of Ljubljana*
Faculty *of Computer and Information Science*

Analysis and comparison of translation errors and biases in LLMs

Pia Polutnik, Tajda Hladnik, Marko Stoklas

Abstract

The abstract goes here. The abstract goes here.

Keywords

Keyword1, Keyword2, Keyword3 ...

Advisors: Slavko Žitnik

Introduction

LLMs have in recent years proven to be quite promising with regard to their capabilities in solving tasks pertaining to natural language processing, including translation. One of the critical areas of concern remains the occurrence of bias embedded within these models, especially in how they handle gendered language during translation. In 2025, extensive research has been conducted that focuses on gender bias in translation. Our project aims to explore this issue further by focusing on translations from English to Slovene. We expect that within these language pairs, which differ in the prevalence of gender markers, the LLM must have a certain means of introducing gender markers. The project investigates what this means by focusing on whether gender biases are introduced and reinforced during the LLM's translation of ambiguous sentences, more specifically how pronouns are translated from English to Slovene, and analyzes to what degree such instances occur. The dataset on which we have decided to conduct our research includes the WinoBias dataset, as we believe it offers a sufficient number of sentences which have already been selected due to them expressing gender bias.

As mentioned, extensive research has been carried out on this matter as of 2025. One such paper – "Investigating Markers and Drivers of Gender Bias in Machine Translations" by Peter J. Barclay and Ashkan Sami (2024) explores the markers of gender bias within machine translation systems. It shows how automated systems propagate traditional gender stereotypes, particularly in the context of professions and roles. The authors identify specific linguistic markers, such as gendered pronouns, that often align with stereotypical gender associations in a certain job, which can inadvertently perpetuate societal biases.

"Biases in Large Language Models: Origins, Inventory, and Discussion" by Roberto Navigli, Simone Conia, and Björn Ross (2024), published in the ACM Journal of Data and Information Quality, also provides a quality examination of the origins of bias in large-scale models. The authors argue that these biases are more than a result of training data, but also stem from the way LLMs learn and process language patterns.

These two research papers underscore the importance of understanding how LLMs handle gendered pronouns, particularly when translating between languages with different gender structures, such as English and Slovene.

Methods

Our project examines gender bias in machine translation from English to Slovene, focusing on occupational bias. The research was conducted using the WinoBias dataset which was accessed through its GitHub repository. A total of 101 sentences were taking from the anti_stereotyped_type2.txt.dev file of the dataset and used in our research. The file contains pairs of sentences where gender stereotypes are intentionally reversed. These sentences are meant to test whether systems translate gender references correctly even when they go against stereotypical expectations. Each sentence follows a specific format: The clerk gave [the mechanic] a present and wished [her] happy birthday. The sentences contain 2 individuals in distinct professions, which we designated as referent 1 (the one that appears first), and referent 2 (the individual mentioned second). One of them (typically referent 1) is not associated with any gendered pronoun, while the other one is. Firstly, we used ChatGPT, a closed-source large language model, to translate the sentences. We fed it the prompt: "Prevedi posamične stavke v slovenščino neodvisno od drugih stavkov." ("Translate individual sentences into Slovene independently from the other sentences"). This prompt ensured that the model processed each sentence in isolation, without context from surrounding text. The dataset was divided among three researchers, with each researcher analysing 34 sentences. This also means that the full set of 102 sentences was not input into ChatGPT all at once; instead, each researcher entered their batch of 34 sentences separately into their own device. The translations were compiled into a Google Docs spreadsheet, where we analysed the results. We looked at the translations of both referents, and whether the Slovene translation rendered them in masculine or feminine forms. We examined how occupations stereotypically linked to a specific gender were translated, noting if the model preserved the anti-stereotypical assignment or reverted to stereotypes. The analysis was conducted by recording how many times each occupation appeared for each of the referents (using the UNIQUE and COUNTIF functions) and recording which gender each referent, as well as the pronoun, was translated to. We also checked whether the genderes of referents matched across the English source sentence and the Slovene translation, and whether the gender assignments were internally consistent within each translation.

Equations

You can write equations inline, e.g. $\cos\pi = -1$, $E = m \cdot c^2$ and α, or you can include them as separate objects. The Bayes's rule is stated mathematically as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{1}$$

where A and B are some events. You can also reference it – the equation 1 describes the Bayes's rule.

Lists

We can insert numbered and bullet lists:

1. First item in the list.
2. Second item in the list.
3. Third item in the list.

- First item in the list.
- Second item in the list.
- Third item in the list.

We can use the description environment to define or describe key terms and phrases.

Word What is a word?.

Concept What is a concept?

Idea What is an idea?

Random text

This text is inserted only to make this template look more like a proper report. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam blandit dictum facilisis. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Interdum et malesuada fames ac ante ipsum primis in faucibus. Etiam convallis tellus velit, quis ornare ipsum aliquam id. Maecenas tempus mauris sit amet libero elementum eleifend. Nulla nunc orci, consectetur non consequat ac, consequat non nisl. Aenean vitae dui nec ex fringilla malesuada. Proin elit libero, faucibus eget neque quis, condimentum laoreet urna. Etiam at nunc quis felis pulvinar dignissim. Phasellus turpis turpis, vestibulum eget imperdiet in, molestie eget neque. Curabitur quis ante sed nunc varius dictum non quis nisl. Donec nec lobortis velit. Ut cursus, libero efficitur dictum imperdiet, odio mi fermentum dui, id vulputate metus velit sit amet risus. Nulla vel volutpat elit. Mauris ex erat, pulvinar ac accumsan sit amet, ultrices sit amet turpis.

Phasellus in ligula nunc. Vivamus sem lorem, malesuada sed pretium quis, varius convallis lectus. Quisque in risus nec lectus lobortis gravida non a sem. Quisque et vestibulum sem, vel mollis dolor. Nullam ante ex, scelerisque ac efficitur vel, rhoncus quis lectus. Pellentesque scelerisque efficitur purus in faucibus. Maecenas vestibulum vulputate nisl sed vestibulum. Nullam varius turpis in hendrerit posuere.

Figures

You can insert figures that span over the whole page, or over just a single column. The first one, Figure 1, is an example of a figure that spans only across one of the two columns in the report.

On the other hand, Figure 2 is an example of a figure that spans across the whole page (across both columns) of the report.

Tables

Use the table environment to insert tables.

Table 1. Table of grades.

Name		
First name	Last Name	Grade
John	Doe	7.5
Jane	Doe	10
Mike	Smith	8

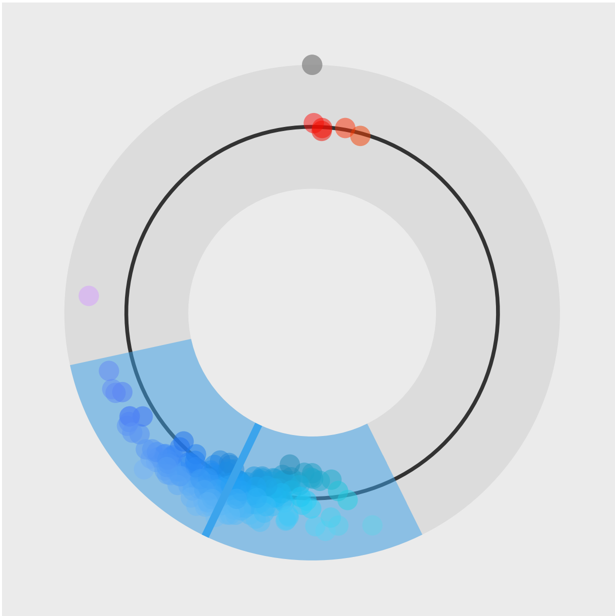

Figure 1. A random visualization. This is an example of a figure that spans only across one of the two columns.

Code examples

You can also insert short code examples. You can specify them manually, or insert a whole file with code. Please avoid inserting long code snippets, advisors will have access to your repositories and can take a look at your code there. If necessary, you can use this technique to insert code (or pseudo code) of short algorithms that are crucial for the understanding of the manuscript.

Listing 1. Insert code directly from a file.

```
import os
import time
import random

fruits = ["apple", "banana", "cherry"]
for x in fruits:
  print(x)
```

Listing 2. Write the code you want to insert.

```
import(dplyr)
import(ggplot)

ggplot(diamonds,
           aes(x=carat, y=price, color=cut)) +
  geom_point() +
  geom_smooth()
```

Results

The analysis you're about to read summarizes the first part of our study – that is how chatGPT contends with translating from English—a language with relatively few grammatical gender markers—into Slovene, which uses gender more explicitly, particularly in pronouns and nouns on account of it being a synthetic language. The analysis draws from the dataset we compiled consisting of 102 English sentences each of which includes one pronoun referring to a professional role, with an even distribution between masculine and feminine pronouns in the source material.

In the source English sentences, there are 51 instances of masculine and feminine pronouns respectively, resulting in a balanced 50/50 gender distribution. The Slovene translations produced by ChatGPT show a slight shift, with 52 masculine and 50 feminine pronouns, yielding a near-balanced, but slightly masculine-leaning, distribution (51% masculine, 49% feminine). These changes are accounted for by two changes from masculine to feminine and one pronoun change from feminine to masculine, which is demonstrated by an increase or decrease in masculine and feminine froms of pronouns in the target language respectively. The occupations that show a shift towards masculine usage are *mechanic* and *laborer,* while the shift towards feminine pronoun from is shown with hairdresser. While the difference is marginal, it indicates a slight tendency toward masculine pronoun usage in the target language, and a gender bias in these three professions.

A deeper analysis was conducted across three scopes of evaluation. The first examined the internal consistency between the gender of the translated pronoun and the gendered form of the profession it refers to—what we refer to as the internal referent-pronoun match. Out of 102 translations, seven sentences (7%) exhibited mismatches where the gender of the pronoun did not align with the gendered form of the referent noun (marked by referent 2). This occurred in professions such as developer, mechanic, housekeeper, mover, chief, and secretary. These mismatches suggest that while the model may correctly translate the pronoun, it does not always ensure that the corresponding profession reflects the same gender, potentially due to underlying gender biases in language modeling. This is interesting since for all other sentences ChatGPT maintained the internal context of the sentence.

The second scope considered the consistency of pronoun gender between the source and target sentences, regardless of internal sentence agreement. In three instances, the translated pronoun in Slovene did not match the intended gender of the original English pronoun, even though there was grammatical agreement within the Slovene sentence. An example of this is demonstrated in the following sentence pair:

"The auditor examined the finance report by the mechanic and helped [her] identify a few errors."

"Revizorka je pregledala finančno poročilo, ki ga je pripravil mehanik, in [mu] pomagala prepoznati nekaj napak."

Here, although "mu" (him) agrees with "mehanik" (mechanic) in Slovene, the original sentence specified a feminine referent ("her"), creating a gender mismatch. This suggests that the model may default to stereotypical gender associations during translation, particularly when the profession is culturally associated with a specific gender, not to mention

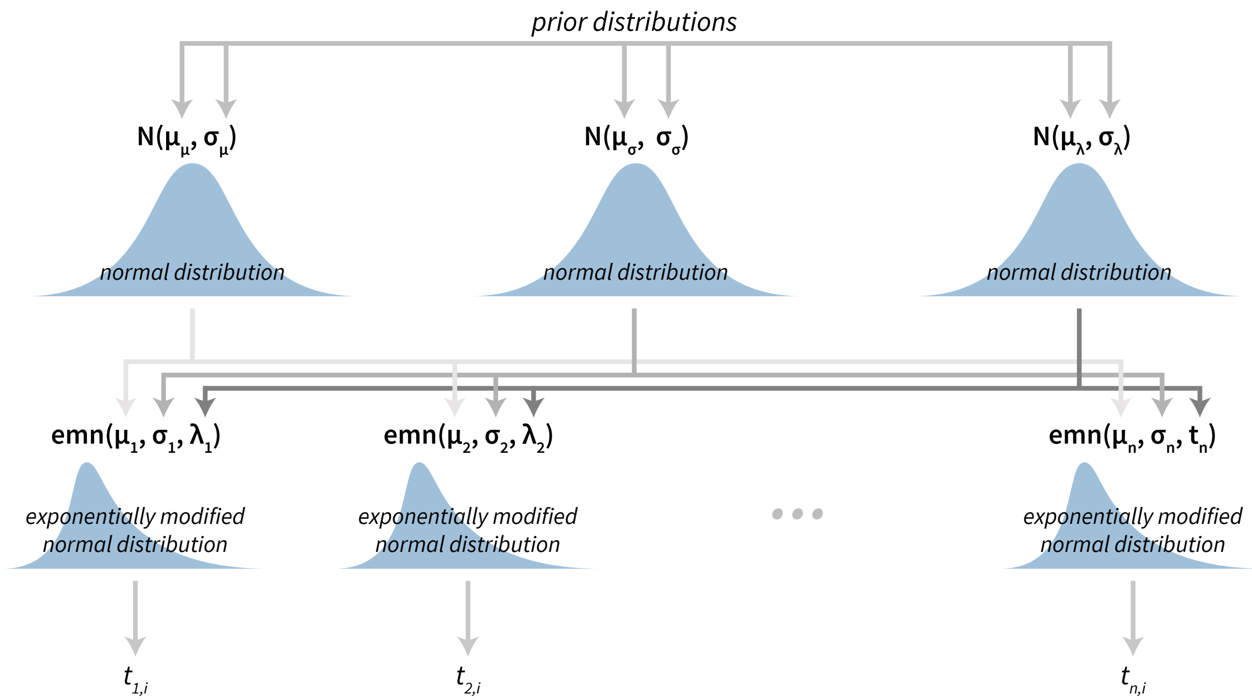

Figure 2. Visualization of a Bayesian hierarchical model. This is an example of a figure that spans the whole width of the report.

that, compared to other sentences, this one is more complex as it contains a dependent relative clause. Mismatches of these nature occur three times, each time with either the profession *mechanic, hairdresser or laborer* – all of which, account also for the unbalanced distribution of pronouns genders within the target text.

Finally, the third analytical scope combined the previous two, assessing whether the gender of the professional referent in the source sentence aligned with the gendered expression of that referent in the target sentence. This metric revealed that 90% of translations (92 out of 102) preserved gender consistency, while 10% failed to do so. This outcome reflects both internal mismatches and pronoun substitutions that collectively indicate areas where the model still struggles to fully preserve gender intentions across languages with differing grammatical systems.

While ChatGPT demonstrates relatively strong performance in translating gendered references from English to Slovene, there remains a consistent, if minor, rate of gender mismatches. These may be accounted for by the model falsely maintaining context when translating sentence, gender bias that occurs in Slovene training data, the tendency of Slovene to use masculine forms as neutral substitutes, or literal translations from source to target languages. More research will be conducted in the following weeks in order to investigate the underlying reason for these phenomena, which may well go beyond gender biases.

As a continuation of this study, we found that, although, largely unimportant in the source language, referent1 becomes a valuable point of research. Although it is not marked by any gendered expression in the source, in Slovene, ChatGPT needs to formulate the noun with a given gender – either masculine or feminine. We took this as an opportunity to study to what degree, if any, the model prioritizes one gender. This is a valuable point of reference for the very questioned proposed above, namely, whether chatGPT uses the masculine form as a neutral stand-in for both genders, or does it show any gender bias. The analysis shows that out of 102 sentences, 77 (76%) show a preference of male referents, while 25 (24%) have the feminine form used for referent 1. Therefore, the masculine form prevails but is hardly the standard – a possible indication of gender bias or the false capturing of context across sentences. The six occupations for which the feminine form of the occupation was opted for referent 1 more than once are: zdravnica (3), varnarka (2), frizerka (2), tajnica (2), kuharica(2), gospodinja (2).

Discussion

Use the Discussion section to objectively evaluate your work, do not just put praise on everything you did, be critical and exposes flaws and weaknesses of your solution. You can also explain what you would do differently if you would be able to start again and what upgrades could be done on the project in the future.

Acknowledgments

Here you can thank other persons (advisors, colleagues ...) that contributed to the successful completion of your project.

References

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in Large Language Models: Origins, Inventory, and Discussion. J. Data and Information Quality 15, 2, Article 10 (June 2023). https://doi.org/10.1145/3597307

Barclay, P. J., & Sami, A. (2024). Investigating Markers and Drivers of Gender Bias in Machine Translations. arXiv.Org, abs/2403.11896. https://doi.org/10.48550/arxiv.2403.11896

Hadas Kotek, Rikker Dockum, David Sun. 2023. Gender bias and stereotypes in Large Language Models. https://dl.acm.org/doi/abs/1