# Realization of Turkish Subject Pronouns in Neural Machine Translation Systems and Large Language Models

**Buket Sak**
University of Konstanz
buket.sak@uni-konstanz.de

**Miriam Butt**
University of Konstanz
miriam.butt@uni-konstanz.de

## Abstract

This study examines how neural machine translation (NMT) systems and Large Language Models (LLMs) handle subject pronoun realization when translating from English into Turkish. English generally disallows the omission of pronouns, whereas Turkish is known as a pro-drop language. One challenge for translation from English to Turkish is thus determining when to drop pronouns and when to realize them overtly. We evaluate NLLB-200, M2M100, OPUS-MT, TranslateGemma, LLaMa 3.1-Instruct, and our fine-tuned variant of LLaMa on a new constructed benchmark data set, which targets subject realization patterns in Turkish, focusing on embedded clauses and contrastive sentences. NLLB-200 and OPUS-MT achieve higher overall translation quality. TranslateGemma aligns more closely with human translations for the pronoun *o* 'he/she', OPUS-MT for the pronoun *sen* 'you', and NLLB-200 for the remaining pronouns. However, the models overall are nowhere near being able to reflect mirror human performance. Our error analysis indicates that a missing understanding of the broader discourse conditions on pronoun realization underlies the current poor performance.

## 1 Introduction

Automatic machine translation has significantly improved with the development of Neural Machine Translation (NMT) and Large Language Models (LLMs). In the context of a larger project studying the ability of LLMs to deal with silent or unexpressed elements of language, we wondered how well LLMs do with languages that allow for *pro-drop*, namely, the omission of pronouns in certain contexts. We present a case study that looks at machine translation (MT) from English, a language that only allows for pro-drop in very restricted contexts (e.g, recipes, diaries; Weir and Haegeman 2026), to Turkish, a language that generally allows for pro-drop (Göksel and Kerslake, 2005).

Most studies have focused on translating from pro-drop to non-pro-drop languages, where the challenge for models is to recover omitted subject pronouns. For example, Guillou (2012) annotates omitted pronouns in the source language for learning in the context of statistical machine translation (SMT) and Wang et al. (2018) argue that subject pronouns are among the most challenging elements to translate in NMT when going from the pro-drop languages Chinese and Japanese to English.

In contrast, relatively little research has examined translation from non-pro-drop to pro-drop languages. In this direction the challenge is to correctly identify when pronouns should or can be dropped and when they should not be dropped. The structural difference between English and Turkish thus serves as a good test case.

Given that there are very few existing resources for English-Turkish MT and none that we know of which focus on pro-drop, our first challenge was to create a data set for benchmarking and evaluation of state-of-the-art NMT and LLMs. To this end, we constructed a data set consisting of embedded clauses and contrastive sentences in English and had these translated into Turkish by two human translators. The intertranslator agreement was high so we could confidently use this data as a benchmark for evaluation. The data focuses particularly on embedded clauses and contrastive sentences because: 1) these were found to be problematic for LLMs in a preliminary pilot study; 2) these contexts allow for both the omission and the overt realization of pronouns — whether or not a pronoun can be dropped depends on a number of linguistic factors, such as co-referentiality (co-referential pronouns should preferably be dropped) and contrast (contrastive contexts generally inhibit pro-drop).

We evaluated the NMT models NLLB-200, OPUS-MT, and M2M100, as well as two LLMs: the translation-specialized TranslateGemma and the general-purpose LLaMa 3.1 Instruct. We also

fine-tuned LLaMa for English–Turkish translation to examine whether this improves subject pronoun use. NMT models generally achieve higher translation quality. Specifically, in contrastive sentences, NLLB-200 achieves the best performance (BLEU: 22.84, chrF: 59.48, COMET: 0.89), while OPUS-MT performs best in embedded clauses (BLEU: 43.56, chrF: 71.28, COMET: 0.94). Fine-tuning LLaMA increases overt subject generation by 69 instances in contrastive sentences (90.14% overt pronoun use) and improves translation quality in embedded clauses (BLEU improves by 15, chrF by 16, and COMET by 0.5). However, all of these results fall far short of human performance and the results are quite poor. We find that while models are able to pro-drop in Turkish, they underperform in contrastive sentences in that these contexts legislate against pro-drop, but the models nevertheless allow for pro-drop. In embedded clauses, pro-drop is connected to the discourse recoverability of the pronoun, a constraint the models are also demonstrably not sensitive to. Overall, our findings contribute to a better understanding of state-of-the-art LLM and NMT models behavior on pronoun realization in the context of translation.

## 2 Background and Related Work

### 2.1 Pronoun Realization

Languages like English allow for the omission of pronouns only in very restricted contexts, such as diaries (*Arrived in Mexico*) or recipes (*Boil for 8 minutes*), e.g., see Weir (2012). It has been well-established that this stands in contrast to other languages, which can either systematically drop only subject pronouns (typical examples are Romance languages, e.g., see Jaeggli (1982)) or all pronouns, e.g., Chinese (Huang, 1984). Turkish is generally classified as a pro-drop language, as it permits the omission of pronouns in a wide variety of contexts (Kornfilt, 1997, pp. 89, 129–130).

In this study we focus on subject pronouns to allow for systematic testing. Göksel and Kerslake (2005, pp. 241–243) note that subject pronouns in Turkish must be preferentially overtly realized in finite clauses when the subject differs from that of the preceding sentence. This is often the case in situations expressing contrast or a topic discontinuity, as shown in (1) with respect to contrast.

(1) Sen uyarı-yı gör-mez-den
you.NOM warning-ACC see-NEG-ABL
gel-ebil-ir-sin ama ben ciddiye
come-PSB-AOR-2SG but I.NOM seriously
al-ıyor-um.
take-PROG-1SG
'You might ignore the warning, but I take it seriously.'

If a subject pronoun were dropped in such contexts, the contrastive interpretation is weaker and the sentence sounds less natural, see (2).

(2) ??Biz her gün çöp-ler-i at-ıyor-uz
we every day trash-PL-ACC throw-PROG-1PL
ama ∅ hiçbir iş yap-mı-yor.
but ∅ no work do-NEG-PROG
'We take out the trash every day, but (he/she) doesn't do any work.'

In embedded clauses we can work with gender contrasts in that in (3-a) the pronouns *he* can refer to either Peter or another referent, but in (3-b) the *she* is taken to refer to somebody other than Peter.

(3) a. Peter$_i$ realized that he$_{i/k}$ had made a mistake.
b. Peter$_i$ realized that she$_{i*/k}$ had made a mistake.

Unlike English, Turkish does not gender its pronouns, but there is a difference in pronominal realization when non-finite clauses are embedded, as in the translation of (3) shown in (4) (Gürel, 2002, p. 29). Here the overt realization of the embedded genitive pronoun *onun* signals that a discourse referent other than Peter made a mistake.

(4) Peter$_i$ onun$_{*i/k}$ bir hata
Peter he/she.GEN a mistake
yap-tığ-ı-nı fark et-ti.
do-NMLZ-3SG-ACC notice do-PST
'Peter$_i$ noticed that he/she$_{*i/k}$ made a mistake.'

In contrast, when the subject of the embedded clause is dropped, the interpretation is left underspecified, since the null subject may refer either to Peter or to another discourse referent. If the pronoun *onun* was dropped in (4), then Peter would be the most likely referent of the null subject, since no other discourse referent is available.

The challenge with both contrastive and embedded clauses for English-Turkish MT is thus to realize/omit pronouns in the correct conditions.

### 2.2 Machine Translation

Machine translation has improved significantly in the last decades, moving from rule-based approaches to machine learning, yielding advanced

models in terms of NMT and LLMs, especially in high-resource settings (Wang et al., 2022; Zhu et al., 2024). Most recent approaches in MT primarily rely on artificial neural networks, as these systems generally produce higher-quality, context-aware translations (Vaswani et al., 2017; Hu, 2024; Wang et al., 2022; Zhang and Zong, 2020).

Building on the transformer architecture, LLMs have also been developed in recent years, typically using a decoder-only architecture and trained on massive text corpora. These models have demonstrated strong performance across various NLP tasks, even without task-specific supervised training (Radford et al., 2019). In the context of MT, general-purpose LLMs can be used, for example, through in-context learning without updating the model's parameters (Brown et al., 2020), or by fine-tuning on specific language pairs. In in-context learning (zero-, one-, and few-shot), models perform tasks using instructions with little or no examples and can achieve strong results without task-specific fine-tuning (Brown et al., 2020).

A major limitation of full fine-tuning is its computational cost. Parameter-Efficient Fine-Tuning (PEFT) methods address this by updating only a small subset of parameters, reducing computational cost while mitigating issues such as overfitting and catastrophic forgetting (Xu et al., 2026). One widely used approach is LoRA (Low-Rank Adaptation) (Hu et al., 2021), which introduces trainable low-rank matrices while keeping the original weights frozen. Its extension, QLoRA (Quantized Low-Rank Adaptation) (Dettmers et al., 2023), further improves efficiency by combining LoRA with quantization, enabling the training of large models on limited hardware. However, these models also present challenges: LLM outputs depend heavily on prompts, may include hallucinations and repetitions, and remain limited for low-resource languages (Ataman et al., 2025).

### 2.3 Machine Translation for Turkish

With respect to pronoun realization in particular, work has shown that NMT systems exhibit great variation in pronoun usage, often overusing (NLLB-600M) or underusing pronouns compared to human translations (Sizov et al., 2024). For the low-resource language Turkish (Acı et al., 2025), previous work by Zhu et al. (2024) found that NLLB (No Language Left Behind) achieves higher BLEU scores than ChatGPT and GPT-4 for Turkish and other Turkic languages. We can also build on some previous existing research that has examined challenges in the English to Turkish translation of pronouns in relation to gender marking and subject omission (Portillo Palma and Alvarez Vidal, 2024; Ciora et al., 2021). This work has shown that both NMT systems and LLMs struggle to fully make use of contextual information, leading to systematic biases and inconsistencies in the translation. While we can build on these studies, they mainly offer insights into gender-related issues in the realization of pronouns, but do not examine broader model performance or pronoun realization (when to omit and when not) in Turkish.

## 3 Methodology

### 3.1 Benchmark Data

Our first task was to construct a data set we could use for evaluation. We were not able to identify existing data that would serve our purposes and therefore constructed our own. Given the known properties of Turkish, cf. §2.1, we decided to focus on two types of sentences: embedded clauses and contrastive sentences.

Embedded clauses were constructed using complementizers (e.g., *that, whether, if*) and interrogative words (*what, who, which, where, why, when*). Each sentence was paired with a version in which the embedded subject either matched or mismatched the matrix subject in gender (e.g., *Maria said that she/he would arrive early*). To control for potential reliance on world knowledge, the data set included both proper names and pronouns. Contrastive sentences involved conjunctions such as *but, though, yet, while, whereas*, and *however*.

In total, we constructed 192 sentences with embedded clauses and 109 contrastive sentences. The numbers are not quite even because the filtering process resulted in different numbers of acceptable instances for each sentence type. We used Perplexity[1] through its chat interface by providing prompts that specified the required linguistic properties of the predefined sentence types to systematically create our benchmark data. The automatically created data was then manually reviewed by the authors and a native English speaker to ensure grammaticality and naturalness, resulting in the final count of 192+109 sentences after the manual weeding out of unacceptable/unnatural sentences, including sentences containing dummy (expletive) pronouns such as *it*.

[1] https://www.perplexity.ai/

All sentences were then translated from English into Turkish by two volunteer human translators, both with bachelor's degrees in translation and a master's degree in linguistics. The embedded clause data set was evenly distributed across translators, and the contrastive data set was approximately balanced (55 vs. 54). To ensure reliability, the translators also evaluated each other's translations using a TQA (Translation Quality Assessment) based on a 5-point Likert scale (with 5 the best score), commonly used to assess MT output (Castilho et al., 2018, p. 10). The human translations both received high scores (Embedded: M = 4.98, SD = 0.14; Contrast: M = 4.91, SD = 0.29), indicating comparable translation quality. The high quality of the translations allowed us to use these as a gold standard reference benchmark for our investigations. [2]

### 3.2 Patterns in the Benchmark Data

**Embedded Clauses** In cases of gender mismatch between the embedded and matrix subjects, human translators consistently used *onun* (third-person singular genitive) in all 96 cases, reflecting overt pronoun realization within the clause. In contrast, when the subjects matched, the embedded subject was never realized overtly. This pattern indicates a very systematic use of omission vs. realization of the pronoun for referential purposes: when the gender mismatches in English the overt pronoun is used; when the gender matches, the pronoun is dropped. That is, although the overt genitive form is not obligatory for referential distinction, it is used to maintain clear referential distinction in embedded clauses.

**Contrastive Sentences** A total of only five out of 213 pronouns were dropped—one each for *he, I, they, we*, and *you*—all by Translator 1. This indicates that pronouns are overwhelmingly expected to be realized overtly in contrastive situations.

### 3.3 Pronoun Identification and Annotation

One crucial step in evaluating the MT performance involves the identification and annotation of the realized or omitted pronouns. To this end, we performed tokenization, POS tagging, and dependency parsing on the generated English data using Stanza (Qi et al., 2020). The corresponding Turkish reference translations were annotated for pronoun realization: dropped pronouns were coded as 0 (omitted) and retained pronouns as 1 (realized). For embedded clauses, we analyzed pronoun drop patterns based on potential coreference, determined by gender agreement: matching gender indicates possible coreference, while mismatch signals non-coreference. We found the automatic annotation to be more challenging for the contrastive translated sentences and ended up using UDPipe 2.3 (Straka, 2018), specifically the pretrained "turkish-penn-ud-2.17-251125" model, to identify the subject pronouns. The annotations were then manually corrected to address missing or incorrectly assigned subject relations. Table 1 shows the distribution of pronouns across both sentence types. Recall that only the third person singular pronouns allow for coreference/non-coreference effects in embedded clauses due to the gender matching in English, which is why we only focused on these third person pronouns for the embedded sentences.

| Pronoun | Embedded | Contrastive |
|---|---|---|
| he | 142 | 42 |
| she | 140 | 40 |
| you | 0 | 46 |
| I | 0 | 45 |
| we | 0 | 20 |
| they | 0 | 20 |

Table 1: Distribution of pronouns across embedded and contrastive sentences.

### 3.4 Machine Translation

We used the following models to investigate how well state-of-the-art NMT and LLMs can handle the translation of pronouns from English into Turkish. Due to computational constraints, models exceeding 8B parameters or closed-source models were not included.

- **opus-tatoeba-en-tr** (Tiedemann, 2020): a bilingual NMT model from the OPUS-MT collection.
- **m2m100-418M** (Fan et al., 2021): a multilingual transformer-based model that enables translation between 100 languages.
- **NLLB-200-3.3B** (Team et al., 2022): a model specifically trained for low-resource languages (although Turkish is not considered a low-resource language in this framework, several other Turkic languages are).

[2] The data and code are available at https://github.com/buketsak/turkish-prodrop-nmt-llm.git

- **TranslateGemma-4B** (Finkelstein et al., 2026): a multilingual MT model based on the Gemma 3 architecture, a decoder-only LLM adapted for translation tasks.
- **LLaMA-3.1-8B-IT** (Grattafiori et al., 2024): an instruction-tuned LLM with 8B parameters designed for multilingual and general-purpose tasks. Turkish is not among the languages officially supported by the model.

### 3.5 Fine-Tuning LLaMa

Since previous work indicated that the fine-tuned LLMs can yield higher translation quality (Zhang et al., 2023; Stap et al., 2024), we also fine-tuned LLaMA-3.1-8B-Instruct for the translation task using QLoRA with 4-bit quantization (BitsAndBytes) to investigate whether performance improvements could be achieved. The model was trained on data from the Tatoeba Project (Tiedemann, 2020), which consists of English–Turkish sentence pairs. The training data were filtered to include sentence structures similar to those in our data set (e.g., explicit personal pronouns, contrastive markers, and similar dependency relations). After filtering, the data set included 25,465 English–Turkish sentence pairs, of which 20,372 were used for training during fine-tuning (setting aside the remainder for potential validation). Training data were formatted in a chat-style structure (system–user–assistant), and LoRA adapters were applied to all attention and feed-forward projection layers. The model was trained for 3 epochs (learning rate $2 \times 10^{-4}$, batch size 1, gradient accumulation 4) on two NVIDIA T4 GPUs (8 h 38 min), without validation due to computational constraints.

Only the first sentence generated by the model was considered to ensure consistent comparison with models producing single-sentence outputs.

Using the same annotation pipeline described in §3.3, embedded clauses were processed with Stanza and contrastive sentences with UDPipe, followed by manual correction. Model outputs were then compared to human translations to assess their alignment with human performance.

## 4 Results and Discussion

### 4.1 Evaluation of Out-of-the-Box Machine Translation Quality

Table 2 shows evaluative scores of the translation quality for each of the MT and LLM models included in this study. The commonly used metrics BLEU (Papineni et al., 2002), chrF (Popović, 2015), and COMET (Rei et al., 2020) were used to assess translation quality, with each metric capturing different aspects of the translated output.

While OPUS-Tatoeba-EN-TR achieves the highest scores across all metrics on embedded clauses (43.56, 71.28, and 0.94, respectively), NLLB-200-3.3B performs best on contrastive sentences (22.84, 59.48, and 0.89, respectively). Overall, the models perform better with respect to the embedded sentences, suggesting there is something about the structure of the contrastive sentences that is challenging for the models.

In the following sections we look at how the models handled subject pronoun realization. To this end, we analyzed both the overall rates of overt pronoun generation and the patterns associated with individual pronouns.

### 4.2 Evaluation of Pronoun Realization

#### 4.2.1 Overt Pronouns in Embedded Clauses

In embedded clauses, subject omission in gender-mismatch cases reflects a choice rather than an error, as omission may introduce ambiguity, whereas overt realization clarifies the referential relations. Given what is known about Turkish and the very consistent usage exhibited by the translators, we expected that models would favor overt subject pronouns in non-coreferential contexts to avoid ambiguity. However, this expectation was not supported, as the models did not produce a sufficient number of overt pronouns to reliably reflect this distinction. In fact, the production of overt pronouns was very low. Of the 96 gender-mismatching instances in which the human translations use an overt pronoun, the models performed as follows: TranslateGemma-4B (29/96, **30.21%**), NLLB-200-3.3B (20/96, 20.83%), OPUS (8/96, 8.33%), LLaMA3.1-IT-8B (5/96, 5.21%), and m2m100-418M (1/96, 1.04%). While none of these models generated third person pronoun *onun* in gender-matched cases, some models generated reflexive pronouns such as *kendi* '-self' and its inflected forms, indicating a preference for maintaining coreference with the subject.

#### 4.2.2 Contrastive Sentences

Our expectation for contrastive sentences was that the models should overwhelmingly realize the subject pronouns overtly, in line with the general pattern observed in Turkish. This expectation was not realized as the models did not match human pro-

| | BLEU | chrF | COMET |
|---|---|---|---|
| | Embedded/Contrast | Embedded/Contrast | Embedded/Contrast |
| OPUS-EN-TR | **43.56**/21.77 | **71.28**/58.15 | **0.94**/0.86 |
| m2m100-418M | 23.85/8.12 | 56.22/46.90 | 0.90/0.78 |
| NLLB-200-3.3B | 39.02/**22.84** | 66.12/**59.48** | 0.93/**0.89** |
| TranslateGemma-4B | 30.30/17.06 | 62.84/54.50 | 0.91/0.88 |
| Llama3.1-IT-8B | 20.92/13.02 | 51.88/49.39 | 0.88/0.82 |

Table 2: Translation quality results for embedded and contrastive sentences, respectively. Best scores in bold.

| Model | Overt | Dropped | O(%) | D(%) |
|---|---|---|---|---|
| human_translations | 208 | 5 | 97.65 | 2.35 |
| opus-tatoeba-en-tr | 154 | 59 | 72.30 | 27.70 |
| m2m100-418M | 44 | 169 | 20.66 | 79.34 |
| NLLB-200-3.3B | **184** | 29 | 86.38 | 13.62 |
| translategemma-4b | 172 | 41 | 80.75 | 19.25 |
| llama3.1-IT-8B | 122 | 91 | 57.28 | 42.72 |

Table 3: Distribution of overt and dropped subjects across models in contrastive sentences (out of 213 cases). O stands for Overt, D for Dropped

| Model | Pronouns | NPs | Total | O(%) | D(%) |
|---|---|---|---|---|---|
| human_translations | 208 | 0 | 208 | 97.65 | 2.35 |
| opus-tatoeba-en-tr | 129 | 25 | 154 | 72.30 | 27.70 |
| m2m100-418M | 44 | 0 | 44 | 20.66 | 79.34 |
| NLLB-200-3.3B | 157 | 27 | **184** | 86.38 | 13.62 |
| translategemma-4b | **169** | 0 | 172 | 80.75 | 19.25 |
| llama3.1-IT-8B | 122 | 0 | 122 | 57.28 | 42.72 |

Table 4: Distribution of overt pronouns, overt noun phrases (NPs), and dropped subjects across translation models in contrastive sentences (out of 213 cases). O stands for Overt, D for Dropped

noun frequencies. This performance thus falls short of human-level accuracy, with model m2m100-418M performing the worst.

Another issue that arose is that when models do not generate the expected pronouns in contrastive sentences, they often produce non-target lexical items. For instance, NLLB-200 frequently generates nouns such as *kadın* 'woman', *erkek* 'boy', *kocası* 'his/her husband', *adam* 'man', and *karısı* 'his/her wife' instead of a pronoun. OPUS-EN-TR exhibits similar behavior with nouns such as *adam* 'man' *kadın* 'woman', and *kız* 'girl' in contexts where the third-person singular pronoun *o* was expected. In other words, when translating *he* or *she*, models—particularly OPUS-EN-TR and NLLB-200-3.3B—tend to replace the neutral pronoun with more specific noun phrases. This tendency may stem from the limited referential specificity of third-person pronouns, which in Turkish are often replaced by explicit noun phrases to ensure clarity (Göksel and Kerslake, 2005, p. 240). While some of these substitutions are acceptable (e.g., using a gendered noun such as *kadın* or *adam*), replacing the subject pronouns with relational noun phrases (e.g., *his wife*, *her husband*) results in incorrect translations. For TranslateGemma, some generated items in subject position, such as *her* 'every' and *diğer* 'other' cannot function as valid subjects in Turkish as they do not constitute a whole NP by themselves.

In our evaluation we thus first examined how often the MT models generated or dropped subjects, regardless of their word class. Table 3 reports the distribution of overt and dropped subjects in subject positions, regardless of noun/pronoun status, while Table 4 provides a breakdown of these subjects by noun phrase and pronoun status. The NLLB model produces the highest number of overt subjects (184), followed by TranslateGemma (172). In contrast, m2m100 shows a clear discrepancy, with the majority of subjects being dropped.

Figure 1 looks at the distribution of only the pronoun realization in more detail in order to assess how human like the models performed with respect to just the targeted pronouns. The left side shows the raw numbers of pronouns generated across the models and the human translators. TranslateGemma-4B generates the most overt pronouns (highest bar after the human translators). Conversely, m2m100-418M produces the least number of overt pronouns (lowest bar). Although NLLB frequently realizes subjects overtly (Table 3), it favors lexical noun phrases over pronouns in third-person contexts, unlike TranslateGemma-4B. This suggests that subject realization and pronoun choice are distinct modeling decisions.

The right side of Figure 1 provides a comparison in terms of percentages of pronouns realized. It can be seen NLLB-200-3.3B matches the human distribution of realized pronouns most closely, despite producing nouns instead of pronouns. In contrast, LLaMA and TranslateGemma tend to produce the

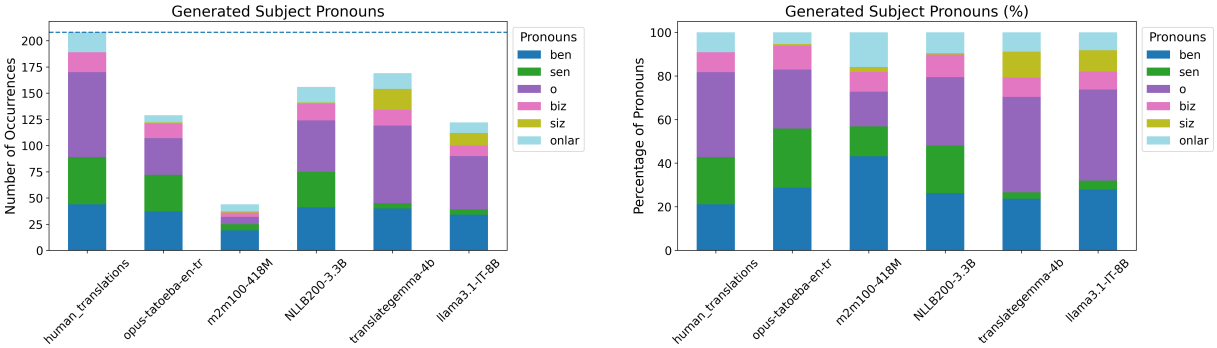


Figure 1: Model vs. Human Translations by Pronoun Occurences and Rates. *ben* (I), *sen* (you), *o* (he/she), *biz* (we), *siz* (you, formal singular or plural), *onlar* (they).

formal form *siz* instead of *sen*, suggesting a bias toward more formal or socially neutral forms for second person pronouns.

In general, NLLB-200 produces the highest number of overt subjects, largely due to its preference for overt third-person singular forms. This partially aligns with Sizov et al. (2024), who report that NLLB-600M generates more pronouns than other models. LLaMA, in contrast, performs poorly, suggesting that the underrepresentation of Turkish in its training data limits translation quality. Nevertheless, it still outperforms m2m100-418M, which may reflect the role of model architecture and scale in shaping the linguistic capabilities of LLMs.

### 4.3 Evaluation of Fine-Tuned LLaMA

Fine-tuning LLaMA led to an improvement in translation quality, particularly for embedded clauses across all metrics (BLEU: 36.11, chrF: 68.65, COMET: 0.93). This corresponds to a gain of 15 BLEU points, 16 chrF points, and 0,5 COMET. In contrast, only modest improvements were observed for contrastive sentences, (BLEU: 13.70, chrF: 50.02, COMET: 0.85).

The fine-tuned model has the highest number of overt genitive pronouns with 30 instances for embedded clauses in mismatching cases, and resulted in having 192 overt subjects (90.14%) and only 21 dropped (9.86%) subjects out of 213 instances in contrastive sentences. Additionally, this model generated one instance of *onun* in a gender-matched case, which was not expected in this context.

Fine-tuning brings the model closer to human performance, and enables it to outperform all other models in our study, even with parameter-efficient training. The gains are particularly pronounced in terms of pronoun realization for contrastive sentences, whereas improvements for embedded clauses remain limited. Overall, the fine-tuned model shows improved handling of referential relations and more appropriate pronoun use.

### 4.4 Error Analysis

In order to understand the models' performance in even more detail, we conducted an error analysis when targeted pronouns were not realized.

**Performance for Embedded Clauses** Some models replace *onun* with non-target genitive forms, including reflexives, demonstratives and lexical nouns. This is particularly evident in TranslateGemma-4B and LLaMA3.1-IT-8B. TranslateGemma-4B also produces lexically appropriate but orthographically incorrect forms such as *o'nun*, as well as pragmatically inappropriate demonstrative genitives like *bunun* 'of this'.

**Performance for Contrastive Sentences** Table 5 reports precision, recall, F1 score, and accuracy for the models on the task of generating overt pronouns in contrastive sentences. Given that pronoun realization is expected in nearly all cases, recall is the most informative metric, as it reflects how effectively models capture contexts requiring pronouns. The fine-tuned model achieves the highest recall (0.91), followed by NLLB-200 (0.88 recall).

Table 6 provides the results for each pronoun. Among the base models, NLLB-200 performs best, with high recall for *ben, biz*, and *onlar*, and the highest overall accuracy (0.74). TranslateGemma achieves the best results for third-person singular *o*, while OPUS performs best for second-person singular forms but shows reduced performance for *o* and *onlar*, likely due to third-person ambiguity. The fine-tuned LLaMA yields the highest scores for all pronouns except *sen*. Overall, NLLB-200, TranslateGemma, and OPUS show relatively strong

| Model | Precision | Recall | F1 | Accuracy |
|---|---|---|---|---|
| opus-en-tr | 1.00 | 0.74 | 0.85 | 0.75 |
| m2m100-418M | 1.00 | 0.21 | 0.35 | 0.23 |
| NLLB-200-3.3B | 1.00 | **0.88** | **0.94** | **0.89** |
| TranslateGemma-4B | 0.98 | 0.81 | 0.89 | 0.80 |
| Llama3.1-IT-8B | 0.99 | 0.58 | 0.73 | 0.58 |
| Fine-tuned Llama3.1-IT-8B | 0.99 | **0.91** | **0.95** | **0.91** |

Table 5: Precision, recall, F1 score, and accuracy for overt subject prediction across models. The top two scores are shown in bold.

| | ben | sen | o | biz | onlar | Accuracy |
|---|---|---|---|---|---|---|
| opus-en-tr | **0.84** | 0.78 | 0.43 | 0.74 | 0.37 | 0.61 |
| m2m100-418M | 0.43 | 0.13 | 0.09 | 0.21 | 0.37 | 0.20 |
| NLLB-200-3.3B | **0.93** | 0.76 | 0.60 | 0.84 | 0.79 | 0.74 |
| TranslateGemma-4B | 0.86 | 0.11 | **0.90** | 0.79 | 0.79 | 0.70 |
| Llama3.1-IT-8B | **0.75** | 0.11 | 0.63 | 0.53 | 0.47 | 0.51 |
| Fine-tuned Llama3.1-IT-8B | 0.93 | 0.69 | 0.96 | 0.95 | **1.00** | 0.89 |

Table 6: Recall and overall accuracy of Turkish pronouns across models. Bold scores indicate the highest score within each model, and underlined scores indicate the highest score across models. The pronoun *siz* is not present as it does not occur in the gold translations. (Full precision, recall, and F1 results are provided in Appendix A.)

performance, whereas LLaMA3.1-IT-8B underperforms without fine-tuning, indicating difficulties in modeling Turkish pronoun realization.

As discussed above and shown in Figure 2, the most prominent pattern is that LLMs tend to produce more formal responses (e.g., using *siz* instead of *sen*). Additionally, NMT systems generate alternative lexical choices instead of *o*, indicating a preference for gender specification.

Moreover, models tend to drop initial subject pronouns while realizing later ones overtly, especially in LLaMA3.1-IT-8B and TranslateGemma-4B, though the pattern also appears in other models. This is unexpected for contrastive sentences, but may be a pattern the models otherwise generally observed in their training data.

Although models may satisfy surface requirements (e.g., overt pronoun use), they may fail to achieve proper structural, morphological, and semantic integration. For example, TranslateGemma-4B exhibits morphosyntactic errors, including incorrect case marking, incorrect suffix realization, and direct lexical transfer from English, such as *bahçelemek* for 'to garden' or *resimlemek* for 'to paint'. LLaMA exhibits similar error patterns, including literal translations that do not correspond to actual words in Turkish (e.g., *danslamak*) along with syntactic, morphological, and case-marking errors. Although the fine-tuned variant produces more subject pronouns, it still shows grammatical limitations, including lexical (e.g., translating 'wallet' as *parmak* 'finger') and morphological errors (e.g., *mutluyorsun* for *mutlu oluyorsun* 'you become happy') and semantic inconsistencies. For example, identical source content was translated differently, resulting in a change in meaning. (hallway → *koridor* in one sentence, but *merdiven* 'stairs') in its paired sentence.

## 5 Conclusion

We evaluated how NMT systems and LLMs handle subject pronoun realization in Turkish, comparing their outputs with human translations in contexts where overt pronouns are required despite the language's overall ability to pro-drop. We found that the models handle embedded subjects well, but tend to drop them in gender mismatch contexts, when they should be overt. Conversely, we also found that although Turkish predominantly allows for pro-drop, the models do produce overt pronouns, suggesting some limited sensitivity to discourse-level constraints.

The findings show that NMT models, particularly NLLB-200 and OPUS-MT, outperformed LLMs in overall translation quality. Results for individual pronouns vary across models, with NLLB-200 showing greater similarity to human translations across a larger number of pronouns. Fine-

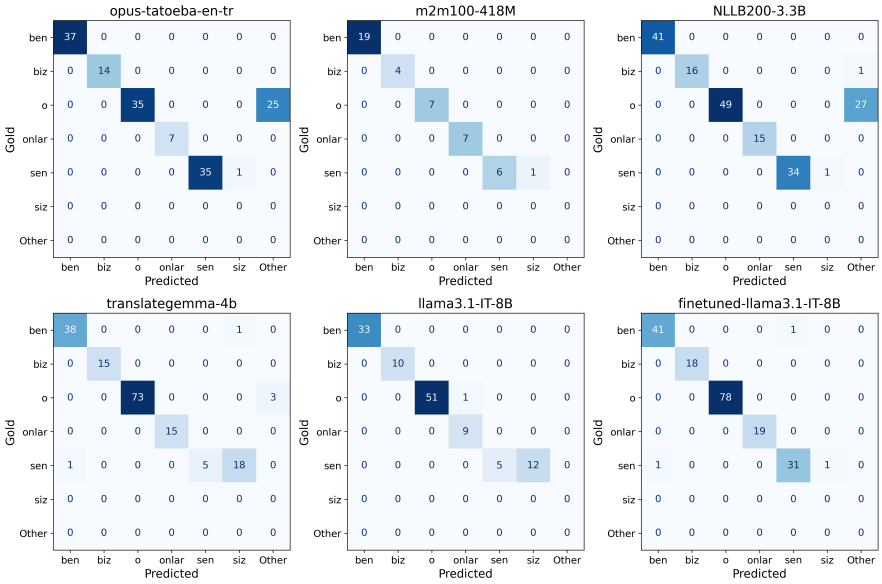


Figure 2: Confusion Matrices of Turkish Personal Pronoun Realizations Across Translation models

tuning improves translation quality for embedded clauses, and increases pronoun realization in contrastive sentences.

Overall, although both NMT systems and LLMs demonstrate some ability to handle subject pronoun realization, they have not yet reached human-level performance and clearly deviate from human translations in Turkish. Even when subjects are present in the source language, the systems fail to generate consistent and reliable output.

The main contributions of this study are as follows: (i) the creation of an annotated data set for conditions governing pro-drop, (ii) the fine-tuning of a LLaMA model for English–Turkish translation using QLoRa, and (iii) an error analysis that reveals linguistic patterns in model outputs.

## Limitations

There are several limitations in our study. First, we did make use of synthetic data to create our benchmark evaluation set. While this was checked by humans, it may still have introduced a bias. Second, we also used names in our embedded clauses, but the association between names and pronouns is not always prototypical, as a given name can correspond to different gender identities (Gautam et al., 2024), adding a level of complexity to the task. Finally, the relatively small size of the data set might limit the generalizability of the findings. Future work should expand the data set, explore additional contexts requiring overt pronouns, and evaluate larger models.

## Acknowledgments

Part of this research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1760, Project A03. We furthermore gratefully acknowledge the resources of the Core Facility LingLab of the Department of Linguistics at the University of Konstanz.

## References

Mehmet Acı, Nisa Vuran Sarı, and Çiğdem İnan Acı. 2025. Morphological and structural complexity analysis of low-resource English-Turkish language pair using neural machine translation models. *PeerJ Computer Science*, 11:e3072.

Duygu Ataman, Alexandra Birch, Nizar Habash, Marcello Federico, Philipp Koehn, and Kyunghyun Cho. 2025. Machine Translation in the Era of Large Language Models: A Survey of Historical and Emerging Problems. *Information*, 16(9).

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, and 10 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Sheila Castilho, Stephen Doherty, Federico Gaspari, and Joss Moorkens. 2018. *Approaches to Human and Machine Translation Quality Assessment*, pages 9–38. Springer International Publishing, Cham.

Chloe Ciora, Nur Iren, and Malihe Alikhani. 2021. Examining covert gender bias: A case study in Turkish and English machine translation models. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 55–63, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond English-Centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, Markus Freitag, and David Vilar. 2026. Translategemma technical report. *Preprint*, arXiv:2601.09012.

Vagrant Gautam, Arjun Subramonian, Anne Lauscher, and Os Keyes. 2024. Stop! in the name of flaws: Disentangling personal names and sociodemographic attributes in NLP. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 323–337, Bangkok, Thailand. Association for Computational Linguistics.

Aslı Göksel and Celia Kerslake. 2005. *Turkish: A Comprehensive Grammar*. Routledge, London.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 540 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Liane Guillou. 2012. Improving pronoun translation for statistical machine translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–10. Association for Computational Linguistics.

Ayşe Gürel. 2002. *Linguistic Characteristics of Second Language Acquisition and First Language Attrition: Turkish Overt versus Null Pronouns*. Ph.D. thesis, McGill University.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Junhui Hu. 2024. Neural Machine Translation (NMT): Deep learning approaches through Neural Network Models. *Applied and Computational Engineering*, 82:93–99.

C.-T. James Huang. 1984. On the distribution and reference of empty pronouns. *Linguistic Inquiry*, 15(4):531–574.

Otto Jaeggli. 1982. *Topics in Romance Syntax*. Foris, Dordrecht.

Jaklin Kornfilt. 1997. *Turkish*. Descriptive Grammars. Routledge.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Şeyda Portillo Palma and Sergi Alvarez Vidal. 2024. Gender bias and contextual sensitivity in machine translation: A focus on Turkish subject-dropped sentences. *transLogos Translation Studies Journal*, 7(2):1–28.

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*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8).

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Fedor Sizov, Cristina España-Bonet, Josef Van Genabith, Roy Xie, and Koel Dutta Chowdhury. 2024. Analysing translation artifacts: A comparative study of LLMs, NMTs, and human translations. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1183–1199, Miami, Florida. Association for Computational Linguistics.

David Stap, Eva Hasler, Bill Byrne, Christof Monz, and Ke Tran. 2024. The fine-tuning paradox: Boosting translation quality without sacrificing LLM abilities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6189–6206, Bangkok, Thailand. Association for Computational Linguistics.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, and 18 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Haifeng Wang, Hua Wu, Zhongjun He, Liang Huang, and Kenneth Ward Church. 2022. Progress in machine translation. *Engineering*, 18:143–153.

Longyue Wang, Zhaopeng Tu, Shuming Shi, Tong Zhang, Yvette Graham, and Qun Liu. 2018. Translating pro-drop languages with reconstruction models. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 4937–4945.

Andrew Weir. 2012. Left-edge deletion in English and subject omission in diaries. *English Language and Linguistics*, 16(1):105–129.

Andrew Weir and Liliane Haegeman. 2026. Register variation: Core grammar and periphery. In Hilary Nesi and Petar Milin, editors, *International Encyclopedia of Language and Linguistics (Third Edition)*, third edition edition, pages 146–154. Elsevier, London.

Lingling Xu, Haoran Xie, S. Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2026. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(6):6107–6126.

JiaJun Zhang and ChengQing Zong. 2020. Neural machine translation: Challenges, progress and future. *Science China Technological Sciences*, 63(10):2028–2050.

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. In *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481, Singapore. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

## A Appendix: Detailed Evaluation Scores

| Model | Pronoun | Precision | Recall | F1 | Overall Acc. |
|---|---|---|---|---|---|
| **OPUS-EN-TR** | *ben* | 1.00 | 0.84 | 0.91 | 0.61 |
| | *sen* | 1.00 | 0.78 | 0.88 | |
| | *o* | 1.00 | 0.43 | 0.60 | |
| | *biz* | 1.00 | 0.74 | 0.85 | |
| | *onlar* | 1.00 | 0.37 | 0.54 | |
| **m2m100-418M** | *ben* | 1.00 | 0.43 | 0.60 | 0.20 |
| | *sen* | 1.00 | 0.13 | 0.24 | |
| | *o* | 1.00 | 0.09 | 0.16 | |
| | *biz* | 1.00 | 0.21 | 0.35 | |
| | *onlar* | 1.00 | 0.37 | 0.54 | |
| **NLLB-200-3.3B** | *ben* | 1.00 | 0.93 | 0.96 | 0.74 |
| | *sen* | 1.00 | 0.76 | 0.86 | |
| | *o* | 1.00 | 0.60 | 0.75 | |
| | *biz* | 1.00 | 0.84 | 0.91 | |
| | *onlar* | 1.00 | 0.79 | 0.88 | |
| **TranslateGemma-4B** | *ben* | 0.97 | 0.86 | 0.92 | 0.70 |
| | *sen* | 1.00 | 0.11 | 0.20 | |
| | *o* | 1.00 | 0.90 | 0.95 | |
| | *biz* | 1.00 | 0.79 | 0.88 | |
| | *onlar* | 1.00 | 0.79 | 0.88 | |
| **Llama3.1-IT-8B** | *ben* | 1.00 | 0.75 | 0.86 | 0.51 |
| | *sen* | 1.00 | 0.11 | 0.20 | |
| | *o* | 1.00 | 0.63 | 0.77 | |
| | *biz* | 1.00 | 0.53 | 0.69 | |
| | *onlar* | 0.90 | 0.47 | 0.62 | |
| **Fine-tuned Llama** | *ben* | 0.98 | 0.93 | 0.95 | 0.89 |
| | *sen* | 0.97 | 0.69 | 0.81 | |
| | *o* | 1.00 | 0.96 | 0.98 | |
| | *biz* | 1.00 | 0.95 | 0.97 | |
| | *onlar* | 1.00 | 1.00 | 1.00 | |

Table 1: Per pronoun precision, recall, and F1 scores for each model.
Overall accuracy is reported once per model.