Do LLMs Recognize *me*, When *I* is not *me*:
Assessment of LLMs Understanding of Turkish Indexical Pronouns in Indexical Shift Contexts

Metehan Oğuz*
USC
moguz@usc.edu

Yusuf Umut Ciftci*
USC
yciftci@usc.edu

Yavuz Faruk Bakman*
USC
ybakman@usc.edu

Abstract

Large language models (LLMs) have shown impressive capabilities in tasks such as machine translation, text summarization, question answering, and solving complex mathematical problems. However, their primary training on data-rich languages like English limits their performance in low-resource languages. This study addresses this gap by focusing on the Indexical Shift problem in Turkish. The Indexical Shift problem involves resolving pronouns in indexical shift contexts, a grammatical challenge not present in high-resource languages like English. We present the first study examining indexical shift in any language, releasing a Turkish dataset specifically designed for this purpose. Our Indexical Shift Dataset consists of 156 multiple-choice questions, each annotated with necessary linguistic details, to evaluate LLMs in a few-shot setting. We evaluate recent multilingual LLMs, including GPT-4, GPT-3.5, Cohere-AYA, Trendyol-LLM, and Turkcell-LLM, using this dataset. Our analysis reveals that even advanced models like GPT-4 struggle with the grammatical nuances of indexical shift in Turkish, achieving only moderate performance. These findings underscore the need for focused research on the grammatical challenges posed by low-resource languages. We released the dataset and code here.

1 Introduction

Large language models demonstrate remarkable capabilities in zero-shot and few-shot learning, excelling across a diverse range of tasks such as machine translation, text summarization, question answering, and solving complex mathematical problems (Ye et al., 2023; OpenAI, 2024; Touvron et al., 2023). However, most large language models (LLMs) are primarily trained on data-rich languages like English, and their performance evaluations are also conducted in these languages (Üstün

et al., 2024). Consequently, this focus on data-rich languages may lead to the under-exploration of challenges unique to low-resource languages.

Recent studies have evaluated the performance of large language models on linguistic tasks such as coreference resolution to examine their ability to match expressions referring to the same entity (Gan et al., 2024; Le and Ritter, 2023; Brown et al., 2020a; Yang et al., 2022; Agrawal et al., 2022). In this study, we investigate LLMs' performance on interpreting indexical pronouns, with a focus on the Indexical Shift problem, a unique linguistic challenge related to but distinct from pronoun resolution, primarily encountered in low-resource languages like Turkish (Şener and Şener, 2011), Amharic (Schlenker, 1999), Zazaki (Anand and Nevins, 2004), Uyghur (Shklovsky and Sudo, 2014), Nez Perce (Deal, 2020) and Japanese (Sudo, 2012).

Indexical elements like *I* and *here* refer to the referents of the speech context such as the speaker or location of utterance. In most languages, these elements must be interpreted within the actual speech context, referring to the actual speaker or location of utterance. However, indexical shift occurs in some languages, like Turkish, where an indexical element can refer to the referents of the reported context, rather than the actual speech context (see Section 2 for details).

Indexical elements are substantially different from pronouns regarding what antecedents they can refer to and what factors restrict their interpretations. For example, while pronouns are almost always ambiguous and can refer to a wide range of entities, first person indexical unambiguously refers to the speaker of the utterance. Even in languages that allow indexical shift, the first person indexical is ambiguous between only two possible referents (the speaker vs the attitude holder), being interpreted based on contex/world knowledge. Moreover, Turkish allows indexical shift

---

*Equal Contribution

only in some syntactic structures (finite embedded clauses) but not in others (e.g. nominalized embedded clauses), which makes indexical shift a unique challenge, requiring attention to specific syntactic rules and context (see Section 3 for a detailed comparison of indexical elements and pronouns regarding coreference resolution).

We investigate the capability of multilingual large language models to handle pronoun resolution within the context of indexical shift in Turkish. To the best of our knowledge, this is the first study examining indexical shift in any language. Therefore, we have released a Turkish dataset specifically designed to evaluate LLMs on the indexical shift problem. Our contributions in this work are as follows:

- We released the Indexical Shift Dataset in Turkish, comprising 156 multiple-choice questions to evaluate LLMs on the indexical shift problem in few-shot setting. Each sample in this dataset includes the necessary linguistic details.

- We evaluate recent multilingual LLMs, including GPT-4, GPT-3.5 (OpenAI, 2024), Cohere-AYA (Üstün et al., 2024), Trendyol-LLM (Trendyol, 2023), and Turkcell-LLM (Turkcell, 2023), using our dataset. We statistically analyze the factors that influence these models' decisions.

- We conclude that even advanced models like GPT-4 struggle to grasp the grammatical nuances of indexical shift in Turkish, showing only moderate performance at best. These findings highlight the need for a special focus on the grammatical challenges of low-resource languages.

2 Indexical Shift in Turkish

Indexical elements. Indexical elements such as English *I, you, here* and *yesterday* are used to refer to referents of the speech-act coordinates (Kaplan, 1977; Schlenker, 2003; Anand and Nevins, 2004; Deal, 2020). For example, *I* is used to refer to *author* (speaker) of the utterance, while *here* is used to refer to the *location* where the utterance was made, and thus sentences like (1) mean different things if uttered by different people and/or in different locations. If (1-a) is uttered by *John* in *Los Angeles*, it means that John was born in Los Ange-

les, but if it is uttered by *Mary* in *Boston* it means that Mary was born in Boston.

(1) a. **I** was born **here**.
     b. Peter thinks that **I** went to Atlanta.

In most languages, including English, indexical elements must always be interpreted inside the actual speech context, referring to actual speech-act coordinates (e.g. author, location).[1] So, if (1-b) is uttered by *John*, the indexical *I* can only be interpreted as referring to John (e.g. John is believed to have gone to Atlanta) but nobody else. Importantly, even though Peter's beliefs are reported in (1-b), the indexical element *I* cannot be interpreted as referring to Peter (author of the reported belief), but must be interpreted as referring to John (author of the actual sentence).

Indexical Shift. Turkish allows indexical shift (e.g. Şener and Şener, 2011), a situation where an indexical element gets its referent from the reported context, rather than the actual context of utterance. For instance, the Turkish first person indexical *ben* in (2) can refer to the attitude holder *Burak*, who is the author of the reported belief, or to the author/speaker of the actual sentence.[2]

(2) Burak yine [(**ben**) mezun ol-du-m   ]
     Burak again   1SG   graduate be-PST-1SG
     san-ıyor.
     think-PROG
     'Burak thinks again that {he/*speaker*} graduated.'

In this regard, sentences like (2) are ambiguous between readings where first person indexical *ben* is shifted (referring to the attitude holder *Burak*) or non-shifted (referring to actual speaker), and Turkish speakers interpret such sentences based on previous context or upcoming sentences (Kuram, 2020). For example, in a context where the actual speaker is the conversation topic, (2) would naturally be interpreted in the non-shifted reading (I = speaker), but if the conversation is about *Burak*, the sentence would be naturally interpreted in the

---

[1]One exception for this generalization is direct quotation (e.g. Peter said/thought, 'I went to Atlanta'.), where quoted material is interpreted as verbatim utterance/thought produced by its owner. Direct quotation is out of the scope of this paper.

[2]Turkish is a *pro*-drop language, meaning that the subject of the clause can be dropped (phonologically null). In this example, and henceforth, parentheses indicate that the subject can optionally be dropped. When subject is dropped, its person features are indicated by the agreement marker on the verb.

shifted reading (I = Burak).[3]

Syntactic Restrictions on Indexical Shift. Indexical shift is a quite rare syntactic/grammatical property, observed in a small set of languages like Amharic (Schlenker, 1999), Zazaki (Anand and Nevins, 2004), Uyghur (Shklovsky and Sudo, 2014), Nez Perce (Deal, 2020). These languages are different from others (e.g. English) in that their syntax possesses the necessary machinery to allow indexical shift (see Deal (2020) for theoretical details and a comprehensive list of languages that allow indexical shift). Moreover, even within a language, indexical shift might be allowed or disallowed depending on the syntactic structure of a sentence. For example, indexical shift in Turkish is observed only with finite embedded clauses like (2), but is not allowed in other grammatical structures such as nominalized embedded clauses like (3), formed by the nominalizer suffix *-DIK* on the embedded verb.

(3) Burak yine [(**ben**-im) mezun
     Burak again 1SG-GEN graduate
     ol-duğ-um-u   ] san-ıyor.
     be-NMLZ-1SG-ACC think-PROG
     'Burak thinks again that {*he/*speaker*}
     graduated.'

Different from (2), the first person indexical *ben* in (3) can only refer to the actual speaker of the sentence (similar to English), regardless of the context it is uttered in (e.g. it cannot undergo indexical shift and refer to the attitude holder Burak). This contrast between (2) and (3) is due to the syntactic/grammatical properties of the finite and nominalized embedded clauses in Turkish, which are acquired by the native speakers of the language (see Şener and Şener (2011) and Oğuz et al. (2020) for syntactic details).

In this study, we aim to test whether LLMs are able to capture this grammatical contrast between two embedded clause types and successfully interpret indexical elements in syntactic environments that allow (e.g. finite embedded clauses) or block indexical shift (e.g. nominalized embedded clauses).

---

[3]Some readers may wonder if embedded material in sentences like (2) is direct quotation (e.g. *Peter thinks, 'I am smart.'*, in English). Özyıldız (2012) and Oğuz et al. (2020) use linguistic diagnostics to show that these are not instances of direct quotation but are true instances of indexical shift.

3   Related Work

There are substantial differences between pronouns and indexical elements. To begin with, even though syntactic/semantic factors can influence pronoun resolution by making some nouns more likely antecedents of pronouns (e.g. subject bias), they do not totally rule out other nouns as possible referents.[4] For example, previous work suggests that speakers mostly interpret the third person pronoun *he* in (4) as referring to the subject *John* (for syntactic or contextual reasons), but the object *Bill* is still a possible antecedent, meaning that the sentence is ambiguous (e.g. Crawley et al., 1990; Stewart and Pickering, 1998; Pickering and Majid, 2007).

(4) John hit Bill and **he** ran away.

Moreover, pronouns can refer to nouns that are contextually salient, but not present in the sentence. For example, the third person pronoun *he* can be interpreted as referring to a contextually salient person named *Peter*. As a result, context plays a crucial role in how speakers interpret pronouns. Previous work in the field (cited above) show that LLMs are able to use contextual information during coreference resolution and show good performance.

Indexical elements, on the other hand, must unambiguously refer to the discourse coordinates (e.g. speaker), except for indexical shifting environments. In syntactic contexts where indexical shift is allowed (e.g. Turkish finite embedded clauses), indexical elements are similar to pronouns in that their referent can be ambiguous. However, indexical elements are still different from pronouns in that they are ambiguous between only two referents (speaker vs attitude holder), while pronouns are technically free to refer to an unlimited amount of antecedents (that can be salient in context).

In summary, indexical elements are restricted by different syntactic factors than pronouns (e.g. clause type) and are usually unambiguous. Moreover, even in contexts where indexical shift is possible, indexical elements are restricted to two possible referents, depending on whether indexical shift takes place or not, while pronouns are free to refer to a wide range of entities. Thus, indexical elements and indexical shift create a unique challenge for LLMs, requiring to take into account the syntactic constraints regarding indexical shift while also

---

[4]Except for ones that violate syntactic principles like the Binding Theory (Chomsky, 1981).

| | | | | Question |
| | | | | NAMENull kimin Almanca bildiğini sanıyor? |
Context	Context prime	Sentence type	Sentence	Ground truth
Merhaba, ben SPEAKER. Ankara'da yaşayan bir öğrenciyim. NAMENull diye bir arkadaşım var. On tane Almanca kelime öğrenmiş. "Hi, my name is SPEAKER. I am a student living in Ankara. I have a friend named NAMENull. He learned ten German words."	Shifted	Finite	NAMENull Almanca biliyorum sanıyor. "NAMEnull thinks he knows German."	Shifted
		Nominalized	NAMENull Almanca bildiğimi sanıyor. "NAMEnull thinks I know German."	Speaker
Merhaba, ben SPEAKER. Ankara'da yaşayan bir öğrenciyim. NAMENull diye bir arkadaşım var. NAMENull söylediklerini havaalanındaki turistler için Almanca'ya çevirmemi istedi. "Hello, I'm SPEAKER. I am a student living in Ankara. I have a friend named NAMEnull. NAMEnull asked me to translate what he said into German for the tourists at the airport."	Speaker	Finite	NAMENull Almanca biliyorum sanıyor. "NAMEnull thinks I know German."	Speaker
		Nominalized	NAMENull Almanca bildiğimi sanıyor. "NAMEnull thinks I know German."	Speaker

Table 1: An example four context-sentence pairs from the dataset.

employing general coreference resolution strategies like contextual information.

4 Turkish Indexical Shift Dataset

To test LLMs' ability to understand indexical shift in Turkish, we created a dataset containing 156 entries with sentences containing the Turkish first person indexical (silent/dropped) that could potentially refer to the speaker (non-shifted) or to the attitude holder of the clause (shifted).

Since the interpretation of indexical elements in Turkish (shifted vs non-shifted) depend on the context they appear in, we created two contexts for each experimental sentence: one priming the shifted reading, the other priming the non-shifted reading of the first-person indexical. Moreover, for each context, we created a version of the sentence with a nominalized embedded clause like (3) (rather than finite embedded clause like (2)), where indexical shift is not allowed by the grammar (e.g. the indexical must refer to the speaker even if the reported context priming otherwise). Together, this four context-sentence pairs lead to four different classes as summarized in Table 1.

We used sentences with three different verbs that allow indexical shift in Turkish: *iste* 'to want', *san* 'to think/believe', *düşün* 'to think'. These verbs trigger specific morphosyntactic requirements in Turkish. For example, *iste* 'to want' requires a subjunctive marker on the embedded verb, while *san* 'to think/believe' and *düşün* 'to think' require regular tense morphology. Also, *düşün* 'to think' requires the complementizer *diye* while the other two does not require/allow *diye*.

Each entry in the dataset have the following information:

- The embedding verb used in the sentence.

- The context first person indexical meaning that the context encourages (context prime).

- Ground truth entity that the indexical is referring to (SPEAKER or SHIFTED).

- The sentence.

- Sentence type (Nominalized or Finite).

- Question to reveal the LLM's interpretation of the indexical.

In order to augment our dataset with different name pairs for testing, SPEAKER and NAMENull placeholders are used instead of the speaker name and the reported third party with NAME-ACC, NAME-GEN, NAME-DAT, NAME-LOC, NAME-COM were used for the accusative, genitive, dative, locative, and comitative forms of the third person subject's name.

The Turkish language indexical shift test dataset is open sourced along with the associated source code, here.

5 Experiments

In this section, we explain our experimental setup, results, and discussion of the results.

5.1 Experimental Design

Models. We assess the performance of five language models: GPT-4 (OpenAI, 2024), GPT-3.5 (Brown et al., 2020b), Cohere-AYA (Üstün et al., 2024), Trendyol-LLM (Trendyol, 2023), and Turkcell-LLM (Turkcell, 2023). Both GPT-4 and GPT-3.5 are closed-source, advanced multilingual models. Cohere-AYA, a 13-billion parameter model, is trained in 101 languages and is built by fine-tuning the mT5 model (Xue et al., 2021). Trendyol-LLM is based on the LLama-2 7-billion model and fine-tuned on both Turkish and English data. Lastly, Turkcell-LLM is a fine-tuned version of the Mistral 7-billion model, specifically adapted for Turkish data.

Evaluation Strategy. We evaluate the performance of models using a multiple-choice question-answer format similar to the Massive Multitask Language Understanding benchmark (MMLU) (Hendrycks et al., 2021) with a 5-shot setting. The

Table 2: Precision, recall, and f1 performances of each model for each class and their macro averages.

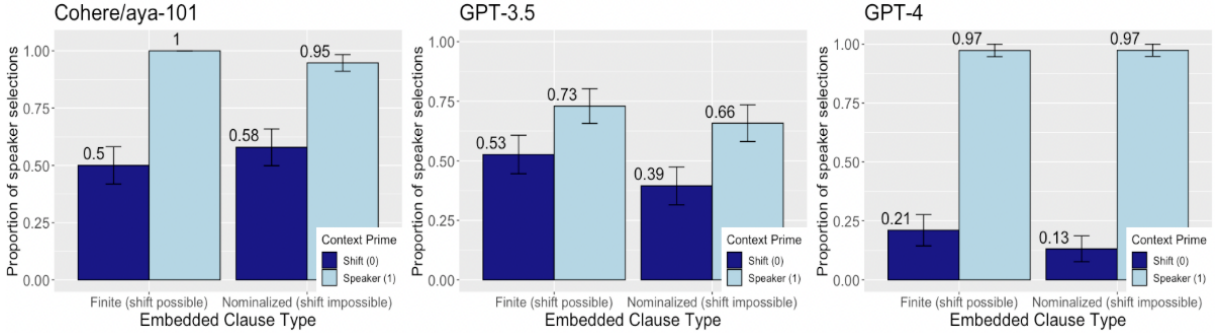| Model	Speaker			Shifted			Macro Average		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
GPT-4	**0.91**	0.69	0.78	0.46	**0.79**	**0.58**	**0.68**	**0.74**	**0.68**
GPT-3.5	0.77	0.59	0.67	0.28	0.47	0.35	0.53	0.53	0.51
Cohere-AYA	0.83	**0.84**	**0.84**	**0.51**	0.50	0.51	**0.67**	0.67	**0.67**
Trendyol-LLM	0.71	0.49	0.58	0.22	0.42	0.29	0.47	0.45	0.43
Turkcell-LLM	0.88	0.12	0.22	0.27	0.95	0.42	0.57	0.54	0.32

Figure 1: Selection analysis plot of GPT-4, GPT-3.5 and Cohere-AYA models. Their outputs are significantly influenced by the context prime, indicating the context's meaning toward either the speaker or the shifted class. No other significant factors were observed.

questions are presented in Turkish, following this template:

```
{in_context_learning_examples}
Soru: {context} {question}?

Seçenekler:
A. {choice_a}
B. {choice_b}

Doğru cevap:
```

where Soru means "question", Seçenekler means "choices", and Doğru cevap means "correct answer". We select five random examples from our dataset as in-context examples and evaluate the remainder. For open-source models, we calculate the probabilities of the tokens "A" and "B" to determine the most likely answer. For closed-source models, we modify the prompt by adding: *"Aşağıdaki soruları cevapla. Sadece cevap olarak A veya B yazman lazım."*[5] to ensure accurate response generation. Both GPT models comply strictly with this rule, generating only the letters A or B as responses.

Our dataset originally contains placeholders such as "NAMENull" and "SPEAKER" instead of real names. We replace these placeholders with

Table 3: Accuracy results for each clause type: finite and nominalized. All models except Cohere-AYA shows worse performance in nominalized sentences.

Model	Finite	Nominalized	Average
GPT-4	**0.88**	0.55	0.72
GPT-3.5	0.60	0.53	0.57
Cohere-AYA	0.75	**0.76**	**0.76**
Trendyol-LLM	0.50	0.50	0.50
Turkcell-LLM	0.57	0.09	0.33

actual random Turkish names to provide more natural linguistic contexts. To decrease the effect of potential gender biases within the models, we uniformly use either female or male names for all placeholders within a single question.

Lastly, to address the choice bias demonstrated in prior studies (Zheng et al., 2024), we implement random assignment of choice options in both the question prompts and in-context examples. Additionally, for open-source models, we enhance reliability by presenting each question twice with the order of choices reversed. We then aggregate the probabilities assigned by the model to each option across these iterations to determine the most likely choice. This method helps to mitigate any inherent preference the model may have towards the position of the answer choices.
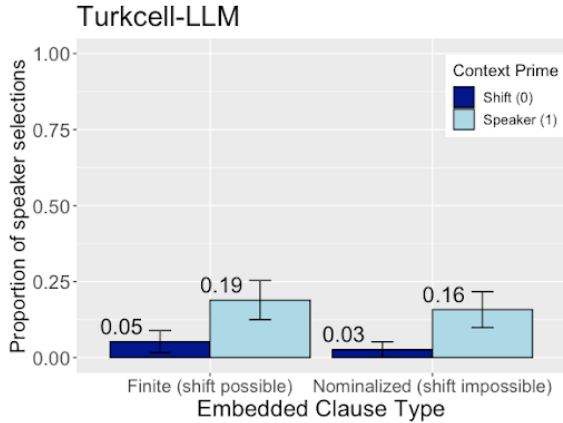
---

[5]In English: Answer the following questions. You need to write only A or B as your answer.

Figure 2: Selection analysis plot of Turkcell-LLM. Neither clause type nor context prime has a statistically significant effect.

Metrics. Our dataset exhibits a class imbalance, with 75% of the ground truth labeled as SPEAKER and the remaining 25% as SHIFTED. Given this imbalance, accuracy alone would not provide a comprehensive measure of model performance; a trivial model that consistently outputs "SPEAKER" could achieve 75% accuracy without truly understanding the data. To address this, we follow the common practice in the literature (Branco et al., 2015) and report precision, recall, and F1 scores for both the SPEAKER and SHIFTED classes. We also compute the macro precision, recall and F1 score to summarize overall performance. Moreover, to see the performance of the models in different clause types (finite or nominalized), we provide the accuracy of all models in different clause types and average accuracy as well. Lastly, analyze factors influencing an LLM's decision-making on a given question, using R Software (R Core Team, 2013) to build the best fitting Linear Mixed-Effect Regression (LMER) model (Bates et al., 2015) with item as the random factor and model selection as the dependent variable. The findings of our statistical analyses are discussed in detail in Section 5.3.

5.2 Main Results

The performance of all models is presented in Table 2. For the Speaker class, GPT-4 achieves the highest precision, while Cohere-AYA consistently delivers high precision and recall, resulting in the highest F1 score for this class. For the Shifted class, GPT-4 attains a maximum F1 score of 0.58, significantly lower than its performance for the Speaker class. All models, except Turkcell-LLM, exhibit lower performance for the Shifted class, indicating a tendency to make mistakes when either the

ground truth or the model output is the class Shifted (low precision and low recall).

Examining the macro average results, we observe that GPT-4 and Cohere-AYA have comparable and highest F1 performance, whereas other models are behind of them with a significant margin. This performance gap for GPT-4 can be attributed to its advanced capabilities, likely due to a large training corpus and model size (OpenAI, 2024). Similarly, the performance gap for Cohere-AYA may be due to its training data, which includes many Turkish samples (Üstün et al., 2024). However, even the performance of GPT-4 and Cohere-AYA is far from optimal. Lastly, as shown in Table 3, GPT-4 and Cohere-AYA perform relatively well at predicting indexical shift in sentences with finite clauses, where shift is possible. However, in sentences with nominalized clauses, where shift is not possible, the performance of all models drops significantly. Among them, only the Cohere-AYA model demonstrates a significantly better prediction accuracy than random guessing (50%). This finding highlights the need for greater attention to the grammatical challenges in low-resource languages.

5.3 Which Factors Effect LLMs' Decision?

In this section, we employ Linear Mixed-Effect Regression (LMER) models to measure the impact of various factors in the dataset on LLM decisions. These factors include sentence type (*finite* vs *nominalized*), gender, and context prime (priming *shifted* vs *non-shifted* readings). Through this analysis, we observe that LLM behaviors can be clustered based on their responses to our task. Below, we examine each LLM cluster and their behavior patterns in detail.

GPT Family and Cohere-AYA. The decisions of these three models are influenced by the context prime, indicating that the context's meaning leans towards either the speaker or shifted class. This effect is statistically significant (p's < 0.001). As illustrated in Figure 1, the models' decisions change significantly when the context prime is altered (represented by dark blue and blue colors). For instance, Cohere-AYA selects the speaker class 100% of the time when the context prime indicates the speaker in finite sentences, but this proportion drops to 50% when the context prime indicates the shifted class. This substantial difference in selection proportions highlights the significant impact
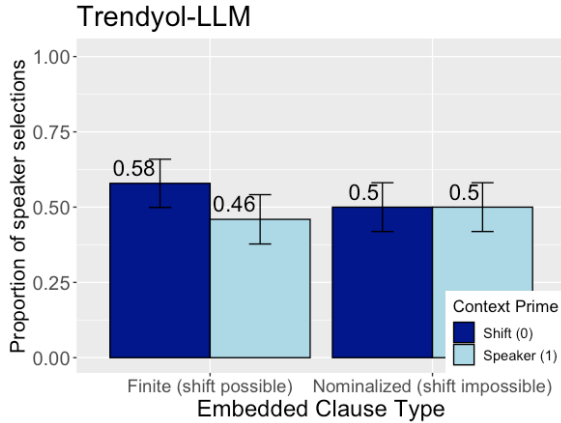
Trendyol-LLM

Figure 3: Selection analysis plot of Trendyol-LLM. Neither clause type nor context prime has a statistically significant effect.

of the context prime, an effect that is also observed in GPT models across different sentence types.

Aside from the effect of context prime, all other factors have a statistically non-significant impact on the models' decisions (p's > .05). This is interesting because, for a native speaker, the clause type (finite or nominalized) directly influences the interpretation of the sentences, allowing indexical shift in finite embedded clauses (2) but not in nominalized embedded clauses (3).

Trendyol-LLM and Turkcell-LLM. Figures 2 and 3 illustrate the mean decisions of these models in each item class. Our analyses show that the decisions of these models are not influenced by either context prime or clause type (p's > .05). Specifically, no factors significantly affect the decisions of Turkcell-LLM, while the only factor that affects the decisions of Trendyol-LLM is interestingly gender (p < .001). Furthermore, Turkcell-LLM almost consistently outputs the shifted class, as seen in Figure 2. Given their comparatively low performance, we interpret these results to mean that these two models lack the reasoning capability to understand the indexical shift problem in Turkish and produce reasonable outputs.

Summary. Overall, none of the models tested in this study are sensitive to clause type, showing that all of these models fail to learn the grammatical grounds where indexical shift is possible (finite embedded clauses) or not (nominalized embedded clauses). Trendyol-LLM and Turkcell-LLM struggle significantly with interpreting the task. The decisions of the other models are primarily affected by the context prime, mimicking native speaker behavior with finite embedded clauses, but over-

generalizing this behavior with nominalized embedded clauses, where context prime does not play a role for native speakers (since indexical shift is not available).

6 Conclusion

In this study, we assess large language models (LLMs) on pronoun resolution tasks within indexical shift contexts, focusing specifically on a low-resource language, Turkish. To facilitate this evaluation, we release a Turkish indexical shift dataset comprising 156 samples. We test recent multilingual models on this dataset and find their performance lacking. Additionally, we observe that none of the LLMs' decisions are influenced by grammatical nuances, such as finite versus nominalized clauses, which contrasts with the behavior of native speakers. Our findings highlight the need for greater attention to the grammatical challenges of low-resource languages in the development and evaluation of LLMs.

7 Limitations

One limitation of the current study is that it concentrated solely on the first person indexical in Turkish, which was due to linguistic limitations regarding indexical shift in Turkish. As explained in detail by Deal (2020), indexical elements within a language do not need to show a uniform behavior, and can have different properties than one another. For example, in Turkish, the person indexicals *ben* 'I' and *sen* 'you' and the temporal indexical *yarın* 'tomorrow' allow indexical shift, while the locative indexical *bura* 'here' does not allow indexical shift (Oğuz et al., 2020). In other words, the locative indexical *bura* 'here' cannot undergo indexical shift, and it must always be interpreted as the location where the sentence was uttered (similar to English *here*). Considering this, the locative indexical *bura* 'here' could not be included in our study. Moreover, the second person indexical *sen* can only shift under the verb *de* 'to say', and cannot shift under other verbs like *san* 'to think' or *iste* 'to want' (because they cannot take an addressee). In the current study, we aimed to observe LLMs' performance under various types of embedding verbs, while keeping our experimental sentences maximally consistent. For this reason, we could not investigate the second person indexical *sen*, which does not allow indexical shift in any other verb than *de* 'to say'. Future work can extend our findings

by investigating LLMs' performance with other indexical elements than the first person *ben*.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David A. Sontag. 2022. Large language models are few-shot clinical information extractors. In *Conference on Empirical Methods in Natural Language Processing*.

Pranav Anand and Andrew Nevins. 2004. Shifty operators in changing contexts. In *Proceedings of SALT XIV*, pages 20–37, Cornell University, Ithaca. CLC Publications.

Douglas Bates, Martin Maechler, Ben Bolker, Steven Walker, Rune Haubo Bojesen Christensen, Henrik Singmann, Bin Dai, and Gabor Grothendieck. 2015. Package 'lme4'. *Convergence*, 12.

Paula Branco, Luis Torgo, and Rita Ribeiro. 2015. A survey of predictive modelling under imbalanced distributions. *Preprint*, arXiv:1505.01658.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Noam Chomsky. 1981. *Lectures on government and binding*. Dordrecht, Holland: Foris.

Rosalind A. Crawley, Rosemary J. Stevenson, and David Kleinman. 1990. The use of heuristic strategies in the interpretation of pronouns. *Journal of Psycholinguistic Research*, 14.

Amy Rose Deal. 2020. *A theory of indexical shift: meaning, grammar, and crosslinguistic variation*. MIT Press, Boston, MA.

Yujian Gan, Massimo Poesio, and Juntao Yu. 2024. Assessing the capabilities of large language models in coreference: An evaluation. In *International Conference on Language Resources and Evaluation*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

David Kaplan. 1977. Demonstratives: An essay on the semantics, logic, metaphysics, and epistemology of demonstratives and other indexicals. *Themes from Kaplan*, pages 565–614.

Kadri Kuram. 2020. L2 Acquisition of the Indexical Shift Parameter in Turkish. *Dilbilim Araştırmaları Dergisi*, 2:231–263.

Nghia T. Le and Alan Ritter. 2023. Are large language models robust coreference resolvers?

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Metehan Oğuz, Burak Öney, and Dennis Ryan Storoshenko. 2020. Obligatory indexical shift in Turkish. In *Proceedings of Canadian Linguistic Association (CLA)*, Western University, London, ON, Canada.

Martin Pickering and Asifa Majid. 2007. What are implicit causality and consequentiality? *Language & Cognitive Processes*, 22.

R Core Team. 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Philippe Schlenker. 1999. *Propositional attitudes and indexicality*. Ph.D. thesis, Massachussetts Institute of Technology.

Philippe Schlenker. 2003. A plea for monsters. *Linguistics and Philosophy*, 26:29–120.

Kirill Shklovsky and Yasutada Sudo. 2014. The syntax of monsters. *Linguistic Inquiry*, 45:381–402.

Andrew J. Stewart and Martin Pickering. 1998. Implicit consequentiality. In *Proceedings of the 20th Annual Conference of the Cognitive Science Society*.

Yasutada Sudo. 2012. *On the Semantics of Phi Features on Pronouns*. Ph.D. thesis, Massachussetts Institute of Technology.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura,

Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Trendyol. 2023. Trendyol LLM 7B Base v0.1. https://huggingface.co/Trendyol/Trendyol-LLM-7b-base-v0.1. Accessed: 2024-05-20.

Turkcell. 2023. Turkcell LLM 7B v1. https://huggingface.co/TURKCELL/Turkcell-LLM-7b-v1. Accessed: 2024-05-20.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. *Preprint*, arXiv:2010.11934.

Xiaohan Yang, Eduardo Peynetti, Vasco Meerman, and Christy Tanner. 2022. What gpt knows about who is who. In *First Workshop on Insights from Negative Results in NLP*.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. A comprehensive capability analysis of GPT-3 and GPT-3.5 series models. *Preprint*, arXiv:2303.10420.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. *Preprint*, arXiv:2309.03882.

Deniz Özyıldız. 2012. When I is not me: A preliminary case study of shifted indexicals in Turkish. Unpublished ms.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *Preprint*, arXiv:2402.07827.

Nilüfer Gültekin Şener and Serkan Şener. 2011. Null subjects and indexicality in Turkish and Uyghur. In *Proceedings of the 7th Workshop on Altaic Formal Linguistics (WAFL7)*, pages 269–284.