TurBLiMP: A Turkish Benchmark of Linguistic Minimal Pairs

Ezgi Başar **Francesca Padovani** **Jaap Jumelet** **Arianna Bisazza**
Center for Language and Cognition (CLCG), University of Groningen
{e.basar, f.padovani, j.w.d.jumelet, a.bisazza}@rug.nl

Abstract

We introduce TurBLiMP, the first Turkish benchmark of linguistic minimal pairs, designed to evaluate the linguistic abilities of monolingual and multilingual language models (LMs). Covering 16 linguistic phenomena with 1000 minimal pairs each, TurBLiMP fills an important gap in linguistic evaluation resources for Turkish. In designing the benchmark, we give extra attention to two properties of Turkish that remain understudied in current syntactic evaluations of LMs, namely word order flexibility and subordination through morphological processes. Our experiments on a wide range of LMs and a newly collected set of human acceptability judgments reveal that even cutting-edge Large LMs still struggle with grammatical phenomena that are not challenging for humans, and may also exhibit different sensitivities to word order and morphological complexity compared to humans.

1 Introduction

A foundational insight in linguistics research is that applying minimal changes to a sentence can render it entirely acceptable or unacceptable to native speakers (Chomsky, 1965). Minimal pairs, as illustrated in Example (1), are a widely used diagnostic tool in linguistics.

(1) a. People in Istanbul love cats.
b. * People in Istanbul loves cats.

Minimal pairs have been a cornerstone of linguistic analysis for decades, and in recent years they have become a vital tool for the linguistic evaluation of language models (LMs). Warstadt et al. (2020) published the first large-scale English **B**enchmark of **Li**nguistic **M**inimal **P**airs (BLiMP) in an effort to systematically evaluate the linguistic knowledge of language models, and since then various benchmarks have been introduced for other languages.

We contribute to this growing collection by introducing the first Turkish benchmark of linguistic minimal pairs. TurBLiMP enriches the typological diversity of available linguistic evaluation benchmarks by incorporating a morphologically rich agglutinative language with highly flexible word order. While Turkish and other agglutinative languages like Finnish have been the object of several studies focusing on word-level morphology (Ismayilzada et al., 2025), the effects of word order flexibility and morphological complexity on the robustness of sentence-level grammatical judgments have not been studied in detail before. We fill this gap by introducing two sets of experimental minimal pair paradigms.

Our evaluation shows that even top-performing LMs suffer performance losses under word order or subordination manipulations, revealing sensitivities that would otherwise go undetected. Compared to the acceptability judgments we collected from native speakers, baseline tests across 13 models and 16 Turkish phenomena demonstrate that Large LMs can struggle with linguistic tasks where humans perform reliably. By providing this resource, we aim to facilitate linguistically motivated NLP research and contribute a high-quality dataset for linguists and NLP researchers.

2 Minimal Pair Benchmarks

Minimal pairs have played an important role for evaluating the linguistic abilities of language models, targeting phenomena such as subject-verb agreement (Linzen et al., 2016), filler-gap dependencies (Wilcox et al., 2018), and negative polarity items (Jumelet and Hupkes, 2018). Warstadt et al. (2020) then established an English benchmark of 67,000 sentence pairs testing 67 paradigms through automated generation based on linguist-curated templates. This work inspired numerous adaptations for other languages, each employing different

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 16507–16522
November 4-9, 2025 ©2025 Association for Computational Linguistics

benchmark creation strategies. Benchmarks using a similar template-based approach as BLiMP include CLiMP (Chinese, Xiang et al., 2021), ZhoBLiMP (Chinese, Liu et al., 2024), BLiMP-NL (Dutch, Suijkerbuijk et al., 2025), and for Basque/Swahili/Hindi by Kryvosheieva and Levy (2025). Another approach is based on modifying Universal Dependency trees, which has been used by SLING (Chinese, Song et al., 2022), RuBLiMP (Russian, Taktasheva et al., 2024), and MultiBLiMP (Jumelet et al., 2025), a multilingual benchmark covering 101 languages. Other approaches include the extraction of minimal pairs from linguistics journals, employed by JBLiMP (Japanese, Someya and Oseki, 2023), manual creation of pairs, as done for Icelandic by Ármannsson et al. (2025), and the usage of LLMs for generating pairs, as done for Tamil and Indonesian by Leong et al. (2023).

Methodological innovations across these benchmarks reveal key trade-offs between scale, linguistic coverage, and data quality. Template-based generation enables large datasets but risks producing unnatural sentences (Vázquez Martínez et al., 2023), while manual extraction from literature or learner corpora ensures quality at the cost of scale. Some of the benchmarks incorporate hybrid approaches and human validation steps to balance these concerns. TurBLiMP too is the result of such hybrid approaches. While creating our benchmark, we developed strategies specifically adapted to the challenges of creating minimal pairs for Turkish.

3 Turkish Morphosyntax & NLP

Turkish presents a particularly interesting case for BLiMP-style evaluation due to its flexible word order and rich morphological system. Turkish syntactically licenses ***all*** **six possible orderings** of the main sentence constituents: Subject-Object-Verb (SOV) represents the canonical order, while other permutations introduce subtle pragmatic variations without altering the core meaning of the sentence. As a result, evaluating LMs on a language like Turkish makes it possible to test them for their robustness to different positional patterns or grammatical hierarchies, in a way that is not possible with English and other fixed-order languages that dominate the training material of current LLMs.

Furthermore, Turkish has highly productive agglutinative morphology, whereby words typically consist of several morphemes attached to a root. Speakers can easily produce and understand numerous legitimate but low-frequency word forms through regular morphological processes, yielding substantially larger vocabulary requirements for LMs compared to analytic and fusional languages. Many syntactic phenomena are realized in Turkish through morphology, rather than by separate function words like in English and other Indo-European languages that form a large chunk of the world's highest-resource languages. A salient example is **subordination**, which largely involves the use of suffixes to nominalize or adverbialize the verb of the embedded clause.

(2) Elif'in Gaye'yi
Elif-3SG.GEN Gaye-ACC
sev-diğ-in-i bil-iyor-um.
like-NMLZ-3SG.POSS-ACC know-PROG-1SG
'I know that Elif likes Gaye.'

For instance, the subordination structure in Example 2 can be intuitively conveyed as 'I know the liking of Gaye by Elif'. Here, the nominalized verb 'like' takes an accusative case suffix as the object of 'know', but also a possessive agreement suffix corresponding to the genitive suffix taken by the subordinate subject 'Elif'.

In general, agglutinative languages such as Finnish, Tamil, Basque, Indonesian or Japanese have been shown to be particularly challenging for neural models (Gerz et al., 2018; Cotterell et al., 2016; Park et al., 2021; Arnett and Bergen, 2025). Focusing on Turkish, Ataman et al. (2017) established that fixed vocabulary constraints combined with suboptimal sub-word segmentation significantly impair neural machine translation performance for agglutinative languages. Ismayilzada et al. (2025) studied LLMs' ability to produce and systematically understand novel well-formed combinations of morphemes in Turkish and Finnish, and reported limited morphological generalization. These findings suggest that studying flexible-order, morphologically rich languages like Turkish can provide unique insights into the true linguistic capabilities of LMs beyond surface fluency.

4 TurBLiMP

The creation of the TurBLiMP benchmark was motivated by the need for a controlled evaluation benchmark that accounts for the unique linguistic properties of Turkish. Some of these properties include flexible word order, morphological richness, optional pro-drop, and syncretism in third-person subject-verb agreement markers. We now provide a brief linguistic background on our minimal pairs.

4.1 Phenomena

We consider 16 different grammatical phenomena, some of which are cross-lingually present in other benchmarks, alongside a few language-specific ones such as suspended affixation (see Table 1 for a complete overview with examples).

ANAPHOR AGREEMENT The anaphoric reflexive pronoun *kendi* agrees with its referent through number and person inflections. Unacceptable sentences in this category feature inflected forms of *kendi* with incorrect agreement.

ARGUMENT STRUCTURE (TRANSITIVE) Turkish has a nominative-accusative case marking system where the direct object of a sentence is marked by the accusative case. However, a special subset of verbs assigns lexical case to their objects, deviating from structural case assignment. Unacceptable sentences feature objects with incorrect case endings, such as dative.

ARGUMENT STRUCTURE (DITRANSITIVE) The prototypical Turkish ditransitive construction applies a dative case marker to the indirect object. However, verbs assigning lexical case can deviate from the general trend. Here too, unacceptable sentences feature objects with incorrect case endings.

BINDING Principle B in Binding Theory (Chomsky, 1981) asserts that pronouns should be free in their binding domain, implying that pronouns should not refer to another entity in the same immediate clause. Unacceptable sentences are created by swapping an anaphora coreferring with the subject with a pronoun of similar features.

DETERMINERS While determiners are largely optional in Turkish, the indefinite article *bir* is sometimes required. When a direct object occurs immediately before the verb, its accusative case ending can be omitted. If such an object is modified by a relative clause, the indefinite article must precede the noun head (Arslan-Kechriotis, 2009). Unacceptable sentences in this phenomenon omit the obligatory determiner.

ELLIPSIS This phenomenon deals with a specific type of ellipsis called backward gapping. For coordinated clauses in Turkish, it is possible to omit the verb in the first clause, leading to a gap which is resolved by the verb in the second clause. Turkish only licenses this if both clauses maintain parallel word order (Bozşahin, 2000). Acceptable sentences show the same subject-object order across clauses while unacceptable ones alternate their order.

IRREGULAR FORMS The aorist is an aspect/ mood marker with three allomorphs -r, -Ir (high vowel harmony), and -Ar (non-high vowel harmony). While monosyllabic verbs take -Ar, a specific subset of irregular verbs take -Ir (Nakipoğlu et al., 2023). Unacceptable sentences feature an incorrect -Ar form.

ISLAND EFFECTS We focus on a specific type of island constraint in which complex noun phrases are modified by a relative clause containing a wh-phrase. The occurrence of the wh-phrase is only permitted if the wh-phrase is *not* an adjunct (Çakır, 2016). Acceptable sentences contain argument wh-phrases like who or what, while unacceptable ones contain wh-adjuncts such as how or why.

NOMINALIZATION Turkish extensively uses a derivational process called nominalization, where verbal bases take suffixes (like -DIK, -mA, and others) to form noun phrases. A category of Turkish verbs only selects complement clauses with -DIK, while others only allow -mA (Kornfilt, 2003b). Correspondingly, minimal pairs contain verbs with the correct and incorrect nominalization suffixes.

NPI LICENSING This phenomenon deals with Turkish negative polarity items such as *hiç*, *kimse*, *hiçbir*, *hiçbir şey*, and *asla*. NPIs occur in contexts where the predicate is negated. Acceptable sentences either omit the NPI or use placeholder indefinite pronouns, while unacceptable ones feature an NPI with a predicate that is not negated.

PASSIVES Turkish licenses the passivization of intransitive verbs via passive suffixes, creating impersonal (vs. personal) passives. While personal passives permit optional by-phrases to express agents, impersonal passives prohibit them (Özsoy, 2009). Thus, acceptable sentences omit by-phrases, while unacceptable ones include them.

QUANTIFIERS Turkish quantifiers such as *her* and *çoğu* can only occur with accusative-marked nouns (Enç, 1991). All minimal pairs for this phenomenon feature direct objects without accusative marking. Unacceptable sentences include a quantifier before the bare noun while acceptable sentences omit it.

Phenomenon	Minimal pair	Translation
Anaphor Agreement	*Gezi rota-sın-ı [kendi-miz /*kendi-niz] internet-e bak-ma-dan oluştur-du-k.* trip route-3SG.POSS-ACC [self-1PL.POSS /*self-2PL.POSS] internet-DAT look-NEG-ABL create-PST-1PL	*We created the trip itinerary [ourselves / *yourselves] without checking the internet.*
Arg. Struct. Trans.	*Eş-im-in [zevk-in-e /*zevk-in-i] çok güven-ir-im.* spouse-1SG.POSS-3SG.GEN [taste-3SG.POSS-DAT /*taste-3SG.POSS-ACC] very trust-AOR-1SG	*I trust my wife's taste a lot.*
Arg. Struct. Ditrans.	*Öğretmen [öğrenci-ler-e /*öğrenci-ler-i] yeni konu-yu anlat-tı.* teacher [student-PL-DAT /*student-PL-ACC] new subject-ACC explain-PST	*The teacher explained the new topic to the students.*
Binding	*Yaz tatil-in-de [kendi-m-i /*ben-i] rahatlamış hissed-iyor-um.* summer holiday-3.POSS-LOC [self-1SG-ACC /*me] relaxed feel-PROG-1SG	*I feel relaxed during the summer holidays.*
Determiners	*Geçen hafta tad-ı damağ-ım-da kal-an [bir /*∅] tatlı ye-di-m.* last week taste-ACC palate-1SG.POSS-LOC stay-PART [a /*∅] dessert eat-PST-1SG	*Last week, I ate a dessert with a taste that lingered on my tongue.*
Ellipsis	*Mağaza-da ceket-i Pelin ve [pantolonu Cem /*Cem pantolonu] seç-ti.* store-LOC jacket-ACC Pelin and trouser-ACC Cem choose-PST	*In the store, Pelin chose the jacket and Cem chose the pants.*
Irregular Forms	*Güneş gör-me-yen petunya-lar hemen [ölür/*öler].* sun see-NEG-PART petunia-PL immediately dies	*Petunias that do not see the sun die immediately.*
Island Effects	*[Neyi /*Onu neden] dükkan-a getir-en eleman azar işit-ti?* [what /*it why] shop-DAT bring-PART worker scolding hear-PST	*The worker who brought what to the store was scolded?*
Nominalization	*Konu-nun tekrar [tartış-ıl-ma-sın-ı /*tartış-ıl-dığ-ın-ı]* öner-iyor-um. matter-GEN again [discuss-PASS-MA-POSS-ACC /*discuss-PASS-DIK-POSS-ACC] suggest-PROG-1SG	*I suggest that the matter be discussed again.*
NPI Licensing	*Kalabalığ-ın ön-ün-de [∅ /*hiç] şarkı söyle-di-m.* crowd-GEN front-POSS-LOC [∅ /*ever] song sing-PST-1SG	*I (*ever) sang in front of a crowd.*
Passives	*Sabah [∅ /*öğrenciler tarafından] okul bahçe-sin-de koş-ul-du.* morning [∅ /*student-PL by] school yard-3SG.POSS-LOC run-PASS-PST	*∼In the morning, it was ran in the school yard (*by the students).*
Quantifiers	*Mağaza-da [∅ /*çoğu] ayakkabı dene-di-m.* store-LOC [∅ /*çoğu] shoe try_on-PST-1SG	*I tried on shoes in the store.*
Relative Clauses	*Sınav-da [gözetmen-in /*gözetmen-i] uyar-dığ-ı öğrenci yer-in-e geç-ti.* exam-LOC [proctor-3SG.GEN /*proctor-ACC] warn-PART-3SG.POSS student place-POSS-DAT move-PST	*The student whom the proctor quietly warned during the exam took his/her seat.*
Scrambling	*Hasan'ın [makale-yi yaz-dığ-ın-ı /*yaz-dığ-ın-ı makale-yi] bil-iyor-um.* Hasan-3SG.GEN article-ACC write-NMLZ-3SG.POSS-ACC know-PROG-1SG	*I know that Hasan wrote the article.*
Subject Agreement	*[Doktor-lar /*Doktor] bu şart-ta çalış-mak zorunda değil-ler.* [doctor-PL /*doctor] this condition-LOC work-NMLZ obliged NEG-3PL	*[Doctors/*Doctor] do not have to work under these conditions.*
Suspended Affixation	*Akşam kız-lar-la parti-ye [git-ti-k /*git] ve çok eğlen-di-k.* evening girl-PL-COM party-DAT [go-PST-1PL /*go] and very have_fun-PST-1PL	*In the evening, we went to a party with the girls and had a lot of fun.*

Table 1: Glossed minimal pairs for each phenomenon in TurBLiMP. The differences are underlined.

Relative Clauses Turkish uses participle suffixes -DIK and -An to form object and subject relative clauses (Göksel and Kerslake, 2005). -DIK clauses feature genitive-possessive agreement. The subject takes genitive case and the verb carries possessive agreement. In subject relative clauses with -An, only the object (if present) is case-marked. Minimal pairs target an argument preceding the nominalized verb. Acceptability depends on whether this noun is inflected with a genitive or non-genitive case ending.

Scrambling Turkish shows word order flexibility and allows postverbal scrambling. This means that constituents can appear after the verb in certain contexts. However, local postverbal scrambling from an embedded clause is prohibited (Kornfilt, 2003a). Acceptable sentences position the object before the embedded verb while unacceptable sentences feature them in the opposite order.

Subject Agreement Turkish realizes subject-verb agreement via person/number suffixes. Gender agreement is absent. A notable feature is third-person syncretism. The same verb inflection can indicate either a third-person singular or plural subject. However, a plural-inflected verb cannot co-occur with a singular subject. Unacceptable sentences either involve singular subjects with plural verbs or pronoun mismatches with first/second-person agreement.

Suspended Affixation Suspended affixation refers to a phenomenon where a shared suffix applies to all conjuncts in a coordinated structure, rather than being repeated. Turkish does not allow suspended affixation for predicates inflected only with the past tense suffix -DI (Serova, 2019). Minimal pairs feature two coordinated past-tense clauses. Acceptable sentences inflect both verbs, while unacceptable ones omit inflection on the first.

4.2 Benchmark Creation

In the creation of TurBLiMP, we opted for the more labor-intensive process of manually crafting sentences. 10 initial samples per each phenomenon were created entirely manually to establish clear guidelines. This first step ensured that each pair differed only minimally while accurately capturing the targeted grammatical contrasts.

Semi-automatic augmentations To enhance lexical diversity, we then adopted a semi-automated workflow in which a masked Turkish LM, BERTurk (Schweter, 2020) is used to suggest lexical replacements at random positions of each manually created sentence. We verified and adjusted each replacement manually to ensure acceptability. This process yielded 100 samples per phenomenon. In a final fully-automated augmentation step, BERTurk was used to generate a list of contextually appropriate words for replacement (e.g. *woman* or *boy* for *girl*). We use the Turkish morphology pipeline by Akın and Akın (2007) to inflect them with the same morphological features when applicable. At the end of this process, our 100 manually validated pairs increase to 1000 pairs per phenomenon. Our three-fold approach balanced scalability with linguistic precision, resulting in a robust benchmark for evaluating Turkish LMs.

4.3 Experimental Paradigms

We further assess the robustness of LMs' syntactic abilities by focusing on two salient properties of Turkish: (i) word order flexibility and (ii) subordination through morphological processes, both discussed in Section 3. Word order variations provide a useful framework for testing the effect of word order biases on syntactic competence, extending the types of variations covered by the existing minimal pair benchmarks (Linzen et al., 2016; Mueller et al., 2020). Subordination is a particularly interesting case to study the interplay between syntactic competence and morphological generalization, broadening the scope of current word-level evaluations (Ismayilzada et al., 2025).

We generate word order and subordinating variations for two of the TurBLiMP phenomena (Transitive and Ditransitive Argument Structure) chosen for their flexibility for manipulation. We derive all 6 subject/verb/object orders and 4 different subordination structures for each minimal pair. Complete examples of experimental paradigms and details about how they were created are provided in Appendix A. The experimental paradigms add a total of 2,000 minimal pairs to the 16,000 pairs forming the base TurBLiMP, and considerably extend our benchmark's utility for investigating controlled linguistic variations.

5 Human Acceptability Judgments

To validate our benchmark, we collected acceptability judgments from 30 native Turkish speakers using a 7-point Likert scale (1: completely unacceptable, 7: completely acceptable). While previous BLiMP variants rely on forced-choice tasks for data validation, BLiMP-NL (Suijkerbuijk et al., 2025) collects Likert scale responses to capture the gradient nature of acceptability judgments. We followed their approach to provide a benchmark that allows for fine-grained evaluation of model-human alignment. Our participant pool was mixed, comprising 17 linguistics students and 13 non-linguists. The study was carried out via an anonymous online survey. Appendix B includes a screenshot of survey instructions. Each participant rated 216 sentences spanning 16 linguistic phenomena as well as 20 experimental paradigms. 3 acceptable and 3 unacceptable sentences were included for each grammatical category, and the acceptability conditions were flipped between the two survey versions.

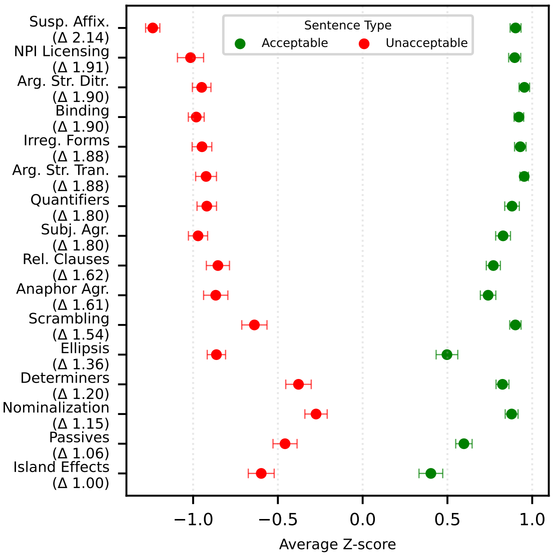


Figure 1: Mean acceptability judgments for 16 TurBLiMP phenomena. Likert scale ratings are transformed to z-scores. Error bars show standard errors of the mean.

Figure 1 reports average acceptability judgments for each phenomenon. Additional participant rating statistics are provided in Appendix C. The responses are first normalized by transforming Likert scores to z-scores. We assume that every participant uses the scale slightly differently, and this step ensures comparability across participants.

Overall, our analysis reveals that participants made clear distinctions between acceptable and unacceptable sentences. The clear separation in z-scores between grammatical and ungrammatical constructions confirms that the targeted syntactic distinctions are perceptible to native speakers. We can also note that some phenomena such as Island Effects, Passives, and Nominalization were less discriminable than others.

Phenomenon	Goldfish 5MB	Goldfish 10MB	Goldfish 100MB	Goldfish 1000MB	BERTurk	cosmosGPT	Gemma 3	Qwen 2.5	Llama 3.1	Aya Expanse	Gemma 2	EuroLLM	Gemma 3
Anaphor Agreement	42.0	44.4	69.9	90.8	97.7	93.0	92.1	83.2	89.5	88.5	89.8	94.1	93.2
Argument Str. Tran.	56.1	50.6	87.9	98.3	99.1	99.3	96.5	82.5	91.3	92.8	92.2	97.6	99.1
Argument Str. Ditr.	65.3	53.2	82.5	92.3	96.1	98.0	96.6	91.7	89.6	94.6	96.8	96.7	97.6
Binding	21.1	25.8	63.3	92.3	99.0	99.2	96.5	91.0	97.1	95.6	97.9	98.6	98.2
Determiners	18.1	25.9	72.5	94.3	99.3	94.2	87.9	75.4	80.1	86.7	91.7	93.3	96.1
Ellipsis	33.0	30.2	68.0	14.9	87.5	40.2	60.5	43.6	63.8	62.7	57.5	73.2	70.4
Irregular Forms	100.0	100.0	100.0	100.0	99.6	100.0	100.0	100.0	100.0	99.4	99.9	100.0	100.0
Island Effects	93.7	92.8	81.0	78.0	51.2	89.8	75.4	86.0	78.3	72.5	75.8	79.3	75.5
Nominalization	66.4	73.9	90.8	93.3	97.4	97.0	94.9	94.0	95.3	92.4	95.9	95.2	96.6
NPI Licensing	93.4	94.3	91.9	98.1	95.0	98.5	97.7	94.0	95.4	96.9	97.1	96.9	98.1
Passives	100.0	100.0	99.9	99.9	81.3	99.4	100.0	100.0	99.6	98.8	99.5	99.7	99.9
Quantifiers	99.0	99.0	99.0	98.9	98.4	99.0	99.0	98.4	98.5	98.0	98.7	98.9	99.0
Relative Clauses	48.3	49.4	76.7	82.0	98.5	93.0	82.4	71.2	80.3	80.6	80.7	83.9	81.5
Scrambling	74.1	86.6	99.6	99.9	100.0	99.7	99.0	99.9	98.2	98.8	100.0	100.0	100.0
Subject Agreement	44.4	41.5	84.1	94.8	98.8	97.5	89.1	82.8	84.8	91.7	90.6	93.7	92.7
Suspended Affixation	57.2	64.8	93.8	98.3	100.0	99.6	99.8	97.5	98.2	98.9	99.5	99.7	100.0
Model Average	63.3	64.5	85.1	89.1	93.7	93.6	91.7	87.0	90.0	90.6	91.5	93.8	93.6
Human Correlation	-0.30	-0.30	0.01	0.16	**0.65**	0.25	0.16	-0.07	-0.01	0.25	0.09	0.17	0.17
Parameter Count	39M	39M	125M	125M	185M	774M	4B	7B	8B	8B	9B	9B	12B
Training Text	Monolingual						Multilingual						

Table 2: Accuracy scores of each model across the linguistic phenomena in TurBLiMP. The red-green color gradient indicates performance, ranging from low to high. Significant Pearson correlations to the human judgments ($p < 0.05$) are indicated in boldface.

6 Experimental Setup

Monolingual models We employed the Goldfish series (Chang et al., 2024), a series of causal LMs with fixed architecture trained on varying training data sizes (5MB, 10MB, 100MB, and 1000MB). Another monolingual model we used is BERTurk (Schweter, 2020), a 185M-parameter Turkish masked LM. With a vocabulary size of 128k, it is the only masked LM in our set of monolingual models. The largest monolingual model that we test is cosmosGPT (Kesgin et al., 2024), a 774M-parameter GPT-2-based model pretrained on Turkish web corpora and books.

Multilingual models The evaluated multilingual models include Qwen 2.5 7B (Yang et al., 2025), Llama 3.1 8B (Meta, 2024), Aya Expanse 8B (Dang et al., 2024), Gemma 2 7B (Team et al., 2024), Gemma 3 4B and 12B (Team et al., 2025), as well as EuroLLM 9B (Martins et al., 2024). For a balanced comparison between the various models, we employed comparable parameter sizes ranging from 4B to 12B. Notably, Aya Expanse is the only instruction-tuned variant in our set of multilingual models, supporting 23 languages including Turkish. The Gemma series also boast multilinguality with Gemma 3 providing support for over 140 languages. EuroLLM prioritizes the coverage of European languages alongside a few others including Turkish.

As our evaluation metric for model performance, we computed entire-sequence log probabilities for acceptable and unacceptable sentences in each pair using the `minicons` library (Misra, 2022; Kauf and Ivanova, 2023). Accuracy scores reflect the proportion of pairs where the model assigned a higher probability to the acceptable sentence. We also report Pearson's correlation between human and model evaluations, calculated from the difference between average scores of acceptable and unacceptable sentences.

7 Results

Model performances across linguistic phenomena are summarized in Table 2. The results reveal that, more often than not, models were able to rate the acceptable sentence higher than its unacceptable counterpart. Some particular phenomena pose challenges for all the models. Ellipsis proved particularly difficult, with scores ranging from 14.9 to 87.5. Other challenging phenomena include Island Effects, Relative Clauses, and Determiners.

Island Effects, Determiners, and Ellipsis also happen to be some of the phenomena with the lowest mean rating difference in acceptability judgments collected from native speakers as seen in Figure 1. We should note that participants preserved a clear acceptability contrast with these phenomena as well. In the case of Ellipsis, considerably low model performances are not consistent with the collected judgments. Though Ellipsis and Scrambling both manipulate word order, models handle Scrambling well. Thus, Ellipsis scores cannot be attributed to general order-manipulation difficulty.

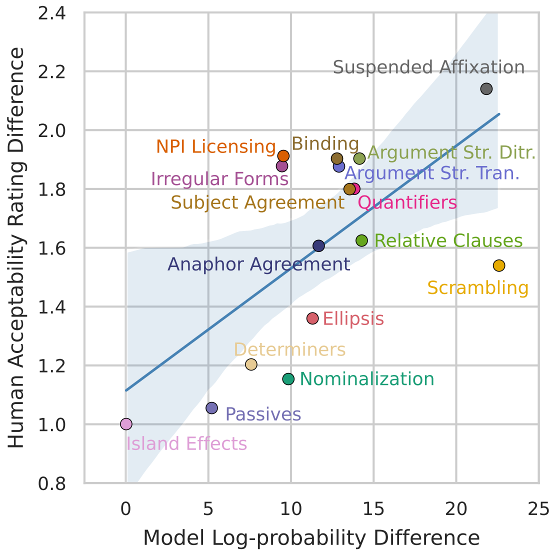


Figure 2: Correlation between the BERTurk model and human acceptability judgments across phenomena. (Pearson's $r = 0.65$, $p = 0.007$) Each data point corresponds to the average difference per phenomenon.

We see that the monolingual models BERTurk and cosmosGPT tend to outperform their multilingual counterparts. Their performance is comparable to the best multilingual models EuroLLM and Gemma 3 12B. BERTurk is the only model that shows a strong cross-phenomenon correlation with human acceptability ratings, as illustrated in Figure 2. This is worth noting given that BERTurk is the only masked language model that we have tested. None of the other models had a statistically significant correlation in either direction. While the reported correlations are based on sequence log probabilities, we also experimented with using Syntactic Log-Odds Ratio (SLOR) (Pauls and Klein, 2012) scores to rule out the possibility that poor correlations were an artifact of the chosen metric. As an acceptability measure, SLOR normalizes sequence probabilities by controlling for sentence length and word frequency. SLOR-based correlations are provided in Appendix D.

Multilingual models generally show better performance with increasing model sizes, but exceptions exist. Gemma 3 4B outperforms Gemma 2 8B, and EuroLLM 9B slightly surpasses Gemma 3 12B. The superior performance of EuroLLM 9B over the same-sized Gemma 2 9B may stem from better distribution of training data across languages.

Finally, the Goldfish model series reveals the effect of training data size on performance. Models with larger training data sizes typically achieve better performance, though some counter-intuitive patterns emerge near random-chance levels. While more data generally improves learning, this pattern does not hold when acceptable sentences are consistently shorter than unacceptable ones. In addition to the reported model results, a supplementary analysis of surface-level confounds such as sentence length and subword counts can be found in Appendix E.

7.1 Effect of Word Order

Our word order paradigm results for the best monolingual (BERTurk) and multilingual (EuroLLM) models are illustrated in Table 3. By manipulating minimal pairs for the Transitive and Ditransitive Argument Structure phenomena, we examine how different word orders affect performance.

	Data	Metric	SOV	SVO	OSV	OVS	VSO	VOS
Human	Slobin and Bever	Freq.	**48.0**	**25.0**	8.0	13.0	6.0	0.0
	BOUN Treebank	Freq.	**59.5**	5.9	4.5	**22.4**	6.8	0.9
	Arg. Str. Tran.	Δ	1.98	1.92	1.81	1.70	1.68	1.75
	Arg. Str. Ditr.	Δ	1.81	1.56	1.47	1.62	1.45	1.29
EuroLLM	Arg. Str. Tran.	Δ	5.24	3.40	2.77	4.39	2.42	2.83
	Arg. Str. Ditr.	Δ	7.83	6.13	5.66	7.11	5.30	4.04
BERTurk	Arg. Str. Tran.	Δ	13.41	10.24	11.92	13.44	7.87	10.16
	Arg. Str. Ditr.	Δ	15.33	8.94	11.21	14.18	7.86	4.74

Table 3: Word order performance comparison between human judgments and best models. The white-blue gradient represents mean acceptability differences (low to high) for each row, while the white-yellow gradient reflects corpus frequency. Frequency values represent the percentage of different word orders within each corpus. For human judgments, mean differences are calculated using z-score–transformed acceptability ratings, whereas for model evaluation, the mean difference reflects the difference in sequence log probabilities.

Although SOV is the canonical word order in Turkish, Slobin and Bever (1982) found that 52% of utterances in their spontaneous adult speech corpus deviate from this order. Similarly, Türk et al. (2022) reported that only 59.5% of sentences in the

BOUN Universal Dependencies Treebank follow SOV. Notably, they identified two different word orders as the second most frequent, highlighting how Turkish word order patterns can vary largely between spoken and written language. Both studies, however, agree that VOS is the least attested.

Native-speaker acceptability judgments reflect that SOV had the highest mean rating difference for both transitive and ditransitive sentences, in line with spoken and written corpus frequencies. The second-highest mean acceptability rating difference for transitive paradigms was the SVO word order, while it was OVS for the ditransitive ones. These are also the second-most-frequent word orders reported by Slobin and Bever (1982) and Türk et al. (2022) respectively. In transitive sentence ratings, VOS is not found to be the most challenging word order, contrary to what might be expected based on attested corpus statistics. This suggests that a rare word order does not inherently hinder people's ability to identify acceptable sentences. Speakers seem to tolerate non-canonical word orders more readily in transitives than in ditransitives. One interpretation may be that case differences are easier to spot in transitive sentences due to fewer arguments.

We see the opposite trend for model evaluations with EuroLLM being particularly sensitive to non-canonical word orders in transitive sentences. BERTurk remains robust to all word orders, showing only a pronounced drop for the rare VOS paradigm in the ditransitive condition. For both transitive and ditransitive sentences, models show high mean log probability differences on OVS word orders. This suggests that model performances align more closely with word order statistics from the BOUN treebank than with those from the spoken language corpus by Slobin and Bever (1982).

7.2 Effect of Subordination

Table 4 displays human and model performance on four subordination paradigms compared to a non-subordinated baseline. In Turkish, subordinate clauses can be finite or non-finite. However, finite subordinate clauses are much less frequent than non-finite ones (Göksel and Kerslake, 2005). For non-finite subordination, we consider three different subordinating suffixes: -DIK, -(y)IncA, and -(y)ken. -DIK forms nominal subordinate clauses while the latter two form adverbial ones.

The acceptability judgment task appears to be easier in non-finite -DIK subordinates than in finite ones, consistent with finite clauses' lower frequency. While -DIK's mean difference nearly matches the baseline in transitives, it shows a decline for ditransitives. Among non-finite structures, -(y)IncA and -(y)ken prove harder than -DIK, suggesting that adverbial clauses pose greater challenges. However, performance deficits may also reflect semantic incongruities from augmentation. Some verb roots may conflict with the aspectual property of the adverbial markers. Therefore, we cannot reliably claim inherent difficulty in adverbial clauses.

		Baseline	Finite	-DIK	-(y)IncA	-(y)ken
Human	Tran. Δ	1.97	1.60	1.95	0.98	0.92
	Ditr. Δ	2.00	1.27	1.59	1.38	1.08
EuroLLM	Tran. Δ	5.34	2.64	4.20	2.97	2.39
	Ditr. Δ	6.83	4.83	5.85	5.31	4.73
BERTurk	Tran. Δ	12.90	8.47	12.10	9.70	10.81
	Ditr. Δ	14.14	9.45	13.43	12.35	11.74

Table 4: Subordination performance comparison between human judgments and best models.

With human judgment patterns established, we evaluate model performance. EuroLLM's -DIK performance shows a drop from baseline in transitive sentences. BERTurk mirrors human trends more closely, exhibiting a greater decline in ditransitives. Both models struggle more with finite subordination than -DIK, though EuroLLM shows a sharper contrast. Compared to nominal subordination, both models show smaller mean differences with adverbial clauses. Overall, we observe that models show sensitivity to different subordination structures.

8 Conclusion

TurBLiMP[1] provides the first comprehensive evaluation of language models' syntactic capabilities for Turkish. We find that larger model sizes generally correlate with higher accuracy, with some exceptions. Considerably smaller monolingual language models often outperform their larger multilingual counterparts and perform on par with the best multilingual models. This finding corroborates patterns attested in other syntactic benchmarks (Taktasheva et al., 2024; Jumelet et al., 2025). The strong performance of monolingual models highlights the importance of language-specific training for reliable models. Cases where smaller models outperformed larger ones also suggest that scaling alone

[1] https://github.com/ezgibasar/turblimp

cannot explain model behavior as far as linguistic evaluations are concerned.

The persistent challenges in phenomena like Ellipsis show that models of all sizes and architectures can struggle with some linguistic phenomena. The discrepancy between model behavior and human judgments for this phenomenon indicates that even the best-performing LLMs may fail to fully capture human linguistic intuition.

TurBLiMP also introduces experimental paradigms to test model robustness to specific linguistic parameters, namely word order and subordination. Results on these paradigms reveal subtle sensitivities in high-performing models that standard evaluations would miss, indicating that even models excelling on general minimal pair tasks can exhibit brittleness with the introduction of non-canonical word orders or subordination. While some performance patterns align with human judgments, we observe both variation across models and cases of divergence from human judgments. Furthermore, the human judgments themselves offer valuable insights into native speaker patterns across different subordination structures and word orders, making TurBLiMP a valuable starting point for future research.

In sum, TurBLiMP provides a valuable resource to assess various linguistic phenomena in a controlled fashion, many of which are not represented in prior syntactic benchmarks. We hope our work will facilitate linguistically informed model developments and contribute to a better understanding of how language models handle linguistic structures across typologically different languages.

Limitations

While TurBLiMP offers a comprehensive evaluation of key linguistic phenomena in Turkish, there are also several limitations to acknowledge. In this paper, we evaluated minimal pair acceptability using sequence log probabilities for each sentence. However, our approach represents only one of several valid methods for assessing language models on acceptability benchmarks. For example, Song et al. (2025) show that prompting for 'metalinguistic' grammaticality judgments can result in better performance than comparing string probabilities directly. However, they do show that this 'introspective' approach has its limitations, and the optimal way of evaluating linguistic ability in LMs remains an open debate (Hu et al., 2024).

Warstadt et al. (2020) define paradigms as minimal pair types and phenomena as broader linguistic categories. Unlike other BLiMP benchmarks that include multiple paradigms under each phenomenon, TurBLiMP currently includes only one paradigm per phenomenon (with the exception of Argument Structure). Splitting some phenomena into multiple paradigms could enable more granular assessment. For a broader coverage of linguistic structures in Turkish, future work could also incorporate additional phenomena into the benchmark.

Our evaluation did not systematically test all available model sizes for each model, which may limit the generalizability of our findings as far as model size is concerned. Testing a wider range of model sizes would strengthen our insights.

Some of the models we tested had both base and instruction-tuned variants. We opted to exclude instruction-tuned versions based on our assumption that English-based tuning was unlikely to improve performance on Turkish grammatical evaluation tasks. Prior work (Chirkova and Nikoulina, 2024) has shown that instruction tuning in English can degrade fluency in non-English languages, and we also observed performance losses when experimenting with the instruction-tuned counterparts of several models. For instance, the average accuracy of Gemma 3 (4B) Instruct was 4.4% lower than its base variant. That said, this experimental choice limits our ability to comment on potential transfer effects from instruction tuning.

Our word order experiments focused exclusively on sentences with explicit subjects, omitting pro-drop constructions. Given Türk et al. (2022)'s finding that subjectless sentences exceed the canonical SOV order frequency in the BOUN treebank, future work should investigate these prevalent but untested configurations.

Acknowledgments

We are thankful to the anonymous reviewers for their helpful comments and to Ümit Atlamaz for his assistance in recruiting participants. We also thank the survey participants for their time and insight. Ezgi Başar received funding from the Erasmus Mundus Masters Programme in Language and Communication Technologies, EU grant no. 2019-1508. Francesca Padovani, Jaap Jumelet and Arianna Bisazza were supported by the Talent Programme of the Dutch Research Council (grant VI.Vidi.221C.009).

References

Ahmet Afşın Akın and Mehmet Dündar Akın. 2007. Zemberek, an open source nlp framework for turkic languages. *Structure*, 10:1–5.

Bjarki Ármannsson, Finnur Ágúst Ingimundarson, and Einar Freyr Sigurðsson. 2025. An Icelandic linguistic benchmark for large language models. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 37–47, Tallinn, Estonia. University of Tartu Library.

Catherine Arnett and Benjamin Bergen. 2025. Why do language models perform worse for morphologically complex languages? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6607–6623, Abu Dhabi, UAE. Association for Computational Linguistics.

Z. Ceyda Arslan-Kechriotis. 2009. Referentiality in turkish: NP/DP. In *Essays on Turkish Linguistics: Proceedings of the 14th International Conference on Turkish Linguistics, 6-8 August 2008*, pages 83–92, Wiesbaden. Harrassowitz Verlag.

Duygu Ataman, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Linguistically motivated vocabulary reduction for neural machine translation from turkish to english. *The Prague Bulletin of Mathematical Linguistics*, 108:331 – 342.

Cem Bozşahin. 2000. Gapping and word order in Turkish. In *Proceedings of the 10th International Conference on Turkish Linguistics*, pages 58–66.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2024. Goldfish: Monolingual language models for 350 languages. *Preprint*.

Nadezhda Chirkova and Vassilina Nikoulina. 2024. Zero-shot cross-lingual transfer in instruction tuning of large language models. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 695–708, Tokyo, Japan. Association for Computational Linguistics.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*, 50 edition. The MIT Press, Cambridge, MA, USA.

Noam Chomsky. 1981. *Lectures on Government and Binding*. Number 9 in Studies in Generative Grammar. Foris Publications, Dordrecht.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared Task—Morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22, Berlin, Germany. Association for Computational Linguistics.

Fred J. Damerau. 1964. A technique for computer detection and correction of spelling errors. *Commun. ACM*, 7(3):171–176.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Mürvet Enç. 1991. The semantics of specificity. *Linguistic Inquiry*, 22(1):1–25.

Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. 2018. On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 316–327, Brussels, Belgium. Association for Computational Linguistics.

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

Omer Goldman, Avi Caciularu, Matan Eyal, Kris Cao, Idan Szpektor, and Reut Tsarfaty. 2024. Unpacking tokenization: Evaluating text compression and its correlation with model performance. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2274–2286, Bangkok, Thailand. Association for Computational Linguistics.

Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, 121(36):e2400917121.

Mete Ismayilzada, Defne Circi, Jonne Sälevä, Hale Sirin, Abdullatif Köksal, Bhuwan Dhingra, Antoine Bosselut, Duygu Ataman, and Lonneke Van Der Plas. 2025. Evaluating morphological compositional generalization in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1270–1305, Albuquerque, New Mexico. Association for Computational Linguistics.

Jaap Jumelet and Dieuwke Hupkes. 2018. Do language models understand anything? on the ability of LSTMs to understand negative polarity items. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 222–231, Brussels, Belgium. Association for Computational Linguistics.

Jaap Jumelet, Leonie Weissweiler, and Arianna Bisazza. 2025. Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs. *Preprint*, arXiv:2504.02768.

Carina Kauf and Anna Ivanova. 2023. A better way to do masked language model scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 925–935, Toronto, Canada. Association for Computational Linguistics.

H. Toprak Kesgin, M. Kaan Yuce, Eren Dogan, M. Egemen Uzun, Atahan Uz, H. Emre Seyrek, Ahmed Zeer, and M. Fatih Amasyali. 2024. Introducing cosmosgpt: Monolingual training for turkish language models. *arXiv preprint arXiv:2404.17336*.

Jaklin Kornfilt. 2003a. *Scrambling, Subscrambling, and Case in Turkish*, chapter 6. John Wiley & Sons, Ltd.

Jaklin Kornfilt. 2003b. *Subject Case in Turkish nominalized clauses*, pages 129–216. De Gruyter Mouton, Berlin, Boston.

Daria Kryvosheieva and Roger Levy. 2025. Controlled evaluation of syntactic knowledge in multilingual language models. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 402–413, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wei Qi Leong, Jian Gang Ngui, Yosephine Susanto, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, and William Chandra Tjhi. 2023. Bhasa: A holistic southeast asian linguistic and cultural evaluation suite for large language models. *Preprint*, arXiv:2309.06085.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Yikang Liu, Yeting Shen, Hongao Zhu, Lilong Xu, Zhiheng Qian, Siyuan Song, Kejia Zhang, Jialong Tang, Pei Zhang, Baosong Yang, Rui Wang, and Hai Hu. 2024. Zhoblimp: a systematic assessment of language models with linguistic minimal pairs in chinese. *Preprint*, arXiv:2411.06096.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. Eurollm: Multilingual language models for europe. *Preprint*, arXiv:2409.16235.

Meta. 2024. Llama 3.1: Open and efficient foundation language models. [Software]. Version 3.1.

Kanishka Misra. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. Cross-linguistic syntactic evaluation of word prediction models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online. Association for Computational Linguistics.

Mine Nakipoğlu, Berna A. Uzundağ, and F. Nihan Ketrez. 2023. Analogy is indispensable but rule is a must: Insights from turkish. *Journal of Child Language*, 50(2):437–463.

Hyunji Hayley Park, Katherine J. Zhang, Coleman Haley, Kenneth Steimel, Han Liu, and Lane Schwartz. 2021. Morphology matters: A multilingual language modeling analysis. *Transactions of the Association for Computational Linguistics*, 9:261–276.

Adam Pauls and Dan Klein. 2012. Large-scale syntactic language modeling with treelets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 959–968, Jeju Island, Korea. Association for Computational Linguistics.

Stefan Schweter. 2020. Berturk - bert models for turkish.

Kutay Serova. 2019. Head movement, suspended affixation, and the turkish clausal spine. *Proceedings of the Workshop on Turkic and Languages in Contact with Turkic*, 4:89.

Dan I. Slobin and Thomas G. Bever. 1982. Children use canonical sentence schemas: A crosslinguistic study of word order and inflections. *Cognition*, 12(3):229–265.

Taiga Someya and Yohei Oseki. 2023. JBLiMP: Japanese benchmark of linguistic minimal pairs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594, Dubrovnik, Croatia. Association for Computational Linguistics.

Siyuan Song, Jennifer Hu, and Kyle Mahowald. 2025. Language models fail to introspect about their knowledge of language. *Preprint*, arXiv:2503.07513.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer. 2022. SLING: Sino linguistic evaluation of large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4606–4634, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Michelle Suijkerbuijk, Zoë Prins, Marianne de Heer Kloots, Willem Zuidema, and Stefan L. Frank. 2025. Blimp-nl: A corpus of dutch minimal pairs and acceptability judgments for language model evaluation. *Computational Linguistics*, pages 1–39.

Ekaterina Taktasheva, Maxim Bazhukov, Kirill Koncha, Alena Fenogenova, Ekaterina Artemova, and Vladislav Mikhailov. 2024. RuBLiMP: Russian benchmark of linguistic minimal pairs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9268–9299, Miami, Florida, USA. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Utku Türk, Furkan Atmaca, Şaziye Betül Özateş, Gözde Berk, Seyyit Talha Bedir, Abdullatif Köksal, Balkız Öztürk Başaran, Tunga Güngör, and Arzucan Özgür. 2022. Resources for turkish dependency parsing: introducing the boun treebank and the boat annotation tool. *Lang. Resour. Eval.*, 56(1):259–307.

Héctor Javier Vázquez Martínez, Annika Heuser, Charles Yang, and Jordan Kodner. 2023. Evaluating neural language models as cognitive models of language acquisition. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 48–64, Singapore. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. CLiMP: A benchmark for Chinese language model evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790, Online. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Sinan Çakır. 2016. Island constraints and adjunct & argument asymmetry in turkish. *Dilbilim Araştırmaları Dergisi*, 27:1–1.

Sumru A. Özsoy. 2009. *Argument structure, animacy, syntax and semantics of passivization in Turkish: A corpus-based approach*, pages 259–279. John Benjamins Publishing Company.

A Experimental Paradigm Details

Our experimental paradigms only target minimal pairs belonging to Argument Structure Transitive and Argument Structure Ditransitive phenomena. Below is an example of the transitive baseline.

(3) a. Bu [şarkı-ya /*şarkı-yı] bayıl-ıyor-um.
this [song-DAT /*song-ACC] love-PROG-1SG
‘I love this song.’

For **word order**, we took the 100 manually crafted minimal pairs from the Transitive and Ditransitive Argument Structure phenomena and, for each phenomenon, we generated 600 pairs for all six possible subject-verb-object permutations. While the original sentences predominantly followed the default Turkish SOV order, the subjects were often omitted and left implicit. We explicitly reintroduced subjects in a dedicated SOV variant and derived the remaining five orders from this augmented set.

(4) a. Ben bu [şarkıya /*şarkıyı] bayılıyorum. (SOV)
b. Ben bayılıyorum bu [şarkıya /*şarkıyı]. (SVO)
c. Bu [şarkıya /*şarkıyı] ben bayılıyorum. (OSV)
d. Bu [şarkıya /*şarkıyı] bayılıyorum ben. (OVS)
e. Bayılıyorum ben bu [şarkıya /*şarkıyı]. (VSO)
f. Bayılıyorum bu [şarkıya /*şarkıyı] ben. (VOS)

	Data	Metric	SOV	SVO	OSV	OVS	VSO	VOS
EuroLLM	Arg. Str. Tran.	Acc.	97.0	91.0	81.0	89.0	82.0	84.0
	Arg. Str. Ditr.	Acc.	98.0	97.0	92.0	94.0	95.0	86.0
BERTurk	Arg. Str. Tran.	Acc.	99.0	98.0	99.0	100.0	96.0	96.0
	Arg. Str. Ditr.	Acc.	98.0	96.0	96.0	98.0	95.0	74.0

Table 5: Accuracy scores for word order paradigms.

For **subordination**, we augmented each paradigm by creating subordinate clauses with three different subordinating suffixes (-(y)IncA, -(y)ken, -DIK) and included a finite subordination paradigm which preserves the original verb inflection. -DIK forms nominal subordinate clauses, while -(y)IncA and -(y)ken form adverbial subordinate clauses. -DIK nominalization carries agreement and case suffixes, and -(y)ken attaches to verb stems marked for aspect.

(5) a. Finite
Bu [şarkıya /*şarkıyı] **bayılıyorum** sanıyor.
'(S)he thinks that I love this song.'

b. -DIK
Bu [şarkıya /*şarkıyı] **bayıldığımı** sanıyor.
'(S)he thinks that I love this song.'

c. -(y)IncA
Bu [şarkıya /*şarkıyı] **bayılınca** gitti.
'∼(S)he left when I really liked this song.'

d. -(y)ken
Bu [şarkıya /*şarkıyı] **bayılırken** gitti.
'∼(S)he left while I was loving this song.'

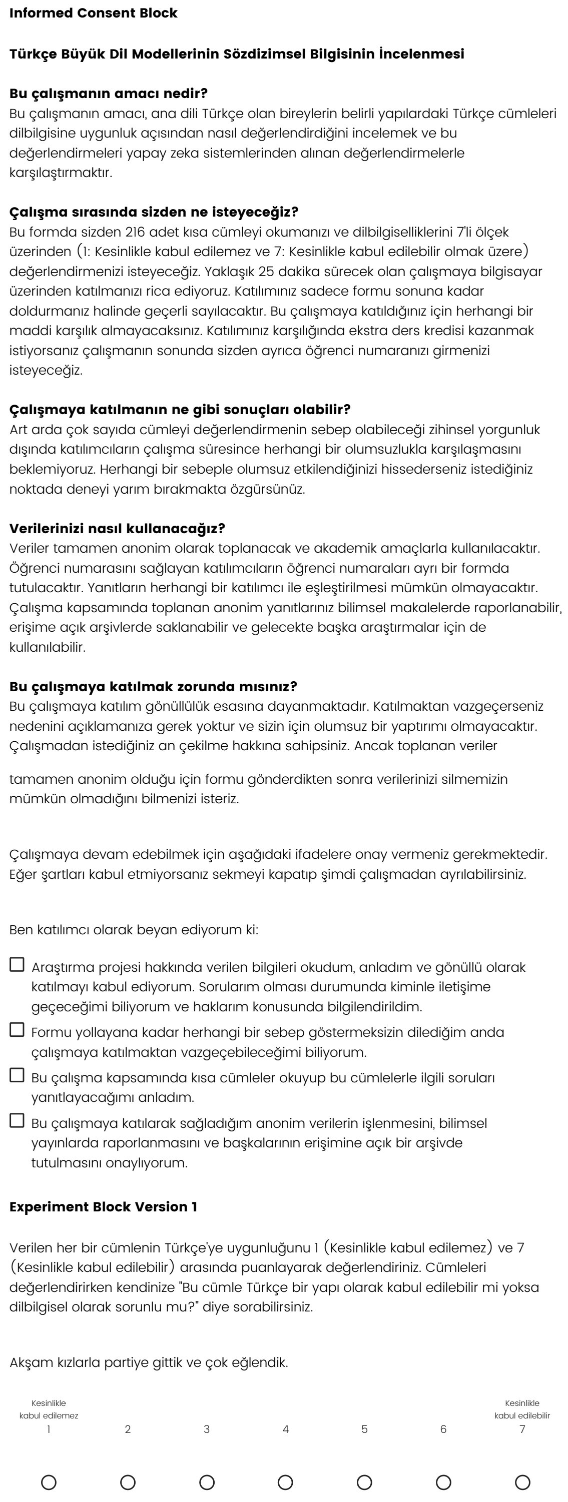
These strategies show varying degrees of morphological complexity, with finite subordination using 2 morphemes, -DIK using 3, -(y)IncA using 1, and -(y)ken using 2. The augmentation procedure yields 400 transitive and 400 ditransitive subordination minimal pairs.

		Baseline	Finite	-DIK	-(y)IncA	-(y)ken
EuroLLM	Tran. Acc.	97.6	82.0	89.0	89.0	77.0
	Ditr. Acc.	96.7	89.0	97.0	94.0	88.0
BERTurk	Tran. Acc.	99.1	97.0	98.0	98.0	97.0
	Ditr. Acc.	96.1	91.0	95.0	96.0	96.0

Table 6: Accuracy scores for subordination paradigms.

B Acceptability Judgment Collection

Figure 3 provides a screenshot of the survey carried out on the Qualtrics platform. All participants gave their informed consent before starting the survey and agreed that their anonymous responses can be made publicly available. Students received extra credit in exchange for their participation. Participation was on a voluntary basis for both students and non-students, and they were explicitly informed that no financial compensation would be provided. The survey was distributed within our close networks.



Informed Consent Block

Türkçe Büyük Dil Modellerinin Sözdizimsel Bilgisinin İncelenmesi

Bu çalışmanın amacı nedir?
Bu çalışmanın amacı, ana dili Türkçe olan bireylerin belirli yapılardaki Türkçe cümleleri dilbilgisine uygunluk açısından nasıl değerlendirdiğini incelemek ve bu değerlendirmeleri yapay zeka sistemlerinden alınan değerlendirmelerle karşılaştırmaktır.

Çalışma sırasında sizden ne isteyeceğiz?
Bu formda sizden 216 adet kısa cümleyi okumanızı ve dilbilgiselliklerini 7'li ölçek üzerinden (1: Kesinlikle kabul edilemez ve 7: Kesinlikle kabul edilebilir olmak üzere) değerlendirmenizi isteyeceğiz. Yaklaşık 25 dakika sürecek olan çalışmaya bilgisayar üzerinden katılmanızı rica ediyoruz. Katılımınız sadece formu sonuna kadar doldurmanız halinde geçerli sayılacaktır. Bu çalışmaya katıldığınız için herhangi bir maddi karşılık almayacaksınız. Katılımınız karşılığında ekstra ders kredisi kazanmak istiyorsanız çalışmanın sonunda sizden ayrıca öğrenci numaranızı girmenizi isteyeceğiz.

Çalışmaya katılmanın ne gibi sonuçları olabilir?
Art arda çok sayıda cümleyi değerlendirmenin sebep olabileceği zihinsel yorgunluk dışında katılımcıların çalışma süresince herhangi bir olumsuzlukla karşılaşmasını beklemiyoruz. Herhangi bir sebeple olumsuz etkilendiğinizi hissederseniz istediğiniz noktada deneyi yarım bırakmakta özgürsünüz.

Verilerinizi nasıl kullanacağız?
Veriler tamamen anonim olarak toplanacak ve akademik amaçlarla kullanılacaktır. Öğrenci numarasını sağlayan katılımcıların öğrenci numaraları ayrı bir formda tutulacaktır. Yanıtların herhangi bir katılımcı ile eşleştirilmesi mümkün olmayacaktır. Çalışma kapsamında toplanan anonim yanıtlarınız bilimsel makalelerde raporlanabilir, erişime açık arşivlerde saklanabilir ve gelecekte başka araştırmalar için de kullanılabilir.

Bu çalışmaya katılmak zorunda mısınız?
Bu çalışmaya katılım gönüllülük esasına dayanmaktadır. Katılmaktan vazgeçerseniz nedenini açıklamanıza gerek yoktur ve sizin için olumsuz bir yaptırımı olmayacaktır. Çalışmadan istediğiniz an çekilme hakkına sahipsiniz. Ancak toplanan veriler

tamamen anonim olduğu için formu gönderdikten sonra verilerinizi silmemizin mümkün olmadığını bilmenizi isteriz.

Çalışmaya devam edebilmek için aşağıdaki ifadelere onay vermeniz gerekmektedir. Eğer şartları kabul etmiyorsanız sekmeyi kapatıp şimdi çalışmadan ayrılabilirsiniz.

Ben katılımcı olarak beyan ediyorum ki:

☐ Araştırma projesi hakkında verilen bilgileri okudum, anladım ve gönüllü olarak katılmayı kabul ediyorum. Sorularım olması durumunda kiminle iletişime geçeceğimi biliyorum ve haklarım konusunda bilgilendirildim.

☐ Formu yollayana kadar herhangi bir sebep göstermeksizin dilediğim anda çalışmaya katılmaktan vazgeçebileceğimi biliyorum.

☐ Bu çalışma kapsamında kısa cümleler okuyup bu cümlelerle ilgili soruları yanıtlayacağımı anladım.

☐ Bu çalışmaya katılarak sağladığım anonim verilerin işlenmesini, bilimsel yayınlarda raporlanmasını ve başkalarının erişimine açık bir arşivde tutulmasını onaylıyorum.

Experiment Block Version 1

Verilen her bir cümlenin Türkçe'ye uygunluğunu 1 (Kesinlikle kabul edilemez) ve 7 (Kesinlikle kabul edilebilir) arasında puanlayarak değerlendiriniz. Cümleleri değerlendirirken kendinize "Bu cümle Türkçe bir yapı olarak kabul edilebilir mi yoksa dilbilgisel olarak sorunlu mu?" diye sorabilirsiniz.

Akşam kızlarla partiye gittik ve çok eğlendik.

Kesinlikle kabul edilemez						Kesinlikle kabul edilebilir
1	2	3	4	5	6	7
○	○	○	○	○	○	○

Figure 3: Informed consent form and instructions.

C Acceptability Judgment Statistics

Category	Proportion
Transitive SOV	30/30
Transitive SVO	30/30
Transitive OSV	30/30
Transitive OVS	29/30
Transitive VSO	30/30
Transitive VOS	30/30
Ditransitive SOV	30/30
Ditransitive SVO	30/30
Ditransitive OSV	28/30
Ditransitive OVS	28/30
Ditransitive VSO	30/30
Ditransitive VOS	29/30

Table 7: Proportion of participants with higher average acceptability ratings for acceptable versus unacceptable sentences per word order paradigm.

Category	Proportion
Transitive Baseline	30/30
Transitive Finite	30/30
Transitive -DIK	30/30
Transitive -(y)IncA	24/30
Transitive -(y)ken	28/30
Ditransitive Baseline	29/30
Ditransitive Finite	29/30
Ditransitive -DIK	30/30
Ditransitive -(y)IncA	29/30
Ditransitive -(y)ken	28/30

Table 8: Proportion of participants with higher average acceptability ratings for acceptable versus unacceptable sentences per subordination paradigm.

Category	Proportion
Anaphor Agreement	30/30
Argument Structure Tran.	30/30
Argument Structure Ditr.	29/30
Binding	30/30
Determiners	28/30
Ellipsis	27/30
Irregular Forms	30/30
Island Effects	26/30
Nominalization	30/30
NPI Licensing	30/30
Passives	29/30
Quantifiers	30/30
Relative Clauses	30/30
Scrambling	30/30
Subject Agreement	30/30
Suspended Affixation	30/30

Table 9: Proportion of participants with higher average acceptability ratings for acceptable versus unacceptable sentences per phenomenon.

D SLOR-based Human Correlations

Model	Pearson's r	p-value
Goldfish 1000MB	0.384	0.1422
BERTurk	**0.557**	**0.0251**
CosmosGPT	0.414	0.1105
Gemma 3 (4B)	0.400	0.1247
Qwen 2.5	0.233	0.3845
Llama 3.1	0.283	0.2881
Aya Expanse	0.454	0.0771
Gemma 2	0.358	0.1730
EuroLLM	0.357	0.1752
Gemma 3 (12B)	0.385	0.1411

Table 10: SLOR-based correlation coefficients for various models and their corresponding p-values. The statistically significant result ($p < 0.05$) is indicated in boldface.

E Phenomenon-Agnostic Factors

To understand general factors influencing model performance beyond phenomena-specific evaluations, we analyzed two potential predictors across all minimal pairs: (1) the difference in sentence lengths between acceptable and unacceptable sentences, and (2) the difference in subword counts for acceptable and unacceptable cue words.

Subword counts refer to the number of smaller units, or subwords, into which a word is segmented by a tokenizer. Depending on the tokenizer design, the resulting subwords may approximate word or morpheme boundaries, but they typically do not perfectly align with either. The subword count difference metric captures the disparity in tokenization granularity between the acceptable and unacceptable cue words and shows whether the acceptable word is split into fewer or more subword units than the unacceptable word.

Using these two predictors, we conduct a linear modeling experiment to predict log probability differences obtained by four different models. We include the two best-performing models (EuroLLM and BERTurk) as well as the two worst-performing models (Goldfish and Qwen 2.5). To allow for better comparability, we opt for the largest Turkish Goldfish model trained on 1000MB of data.

Figure 4 displays the standardized β coefficients for both sentence length differences and subword count differences across all four models. We also include each model's corresponding R^2 value, indicating how well these predictors explain the variance in log probability differences.

R^2 values are remarkably low for all the models, which means that these two factors alone are not adequate predictors of model behavior. For the purposes of our benchmark, this is the desired outcome as we want model log probabilities to reflect grammatical judgments rather than being confounded by these surface-level features.

Sentence length Autoregressive and masked models show sentence length effects in opposite directions. In the case of autoregressive models, the model is more likely to incorrectly prefer the unacceptable sentence if the acceptable sentence is longer than its unacceptable counterpart. This indicates that autoregressive models are biased to prefer shorter sequences. BERTurk, which is the only masked language model in our model pool, shows the opposite pattern. As the acceptable sentence gets longer than the unacceptable one, BERTurk is more likely to make the correct prediction. This is possibly due to longer sentences providing more contextual clues for masked token prediction.

Reflecting back on the previously reported accuracy results, we can note that acceptable sentences are typically shorter than unacceptable ones in NPI Licensing, Passives and Quantifiers while the opposite is true in Determiners. Autoregressive models perform better on NPI Licensing, Passives and Quantifiers while BERTurk exhibits slightly better performance for Determiners. This behavioral pattern aligns with the results of our linear model.

Subword counts Based on prior work (Goldman et al., 2024; Jumelet et al., 2025), we expect model performance to degrade with increasing subword count differences. The Goldfish model shows a negative subword count coefficient, behaving in line with our hypothesis. A negative coefficient indicates that the model is more likely to make the wrong prediction if the acceptable cue word is split into more subwords than the unacceptable one. However, contrary to our expectations, all other models show a weak but consistent positive effect. To help us make sense of this counterintuitive finding, we conduct a follow-up analysis examining how each model's tokenizer aligns with morphological boundaries in our stimuli.

Utilizing the morphological inflection pipeline by Akın and Akın (2007), we construct a small dataset of 400 morphologically segmented Turkish words. The dataset is organized into four categories,

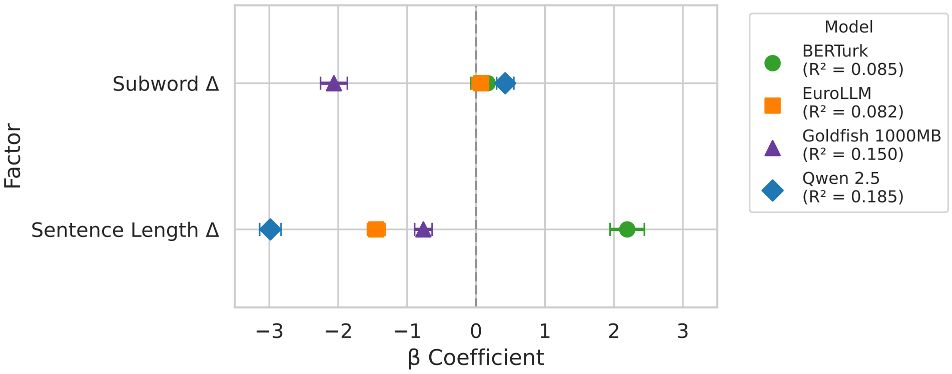


Figure 4: β coefficients fitted for the BERTurk, EuroLLM, Goldfish, and Qwen 2.5 models with sentence length and subword count differences as the predictors.

each targeting morphemes that frequently appear as part of the cue words in our benchmark.

- **(Verb + Nominalizer + Possessive + Case)**: Contains 100 verb forms with the nominalizer *-DIK* followed by possessive markers and case endings (e.g., *unuttuğunu* → [`unut`, `tuğ`, `un`, `u`]). This category evaluates the segmentation of nominalized verbs.
- **(Verb + TAM + Person)**: Includes 100 finite verbs segmented into stems, tense/aspect/modality (TAM) markers, and person markers (e.g., *olacaktılar* → [`ol`, `acak`, `tı`, `lar`]), targeting inflectional morphology.
- **(Noun + (Plural) + Possessive + Case)**: Comprises 100 possessed nouns decomposed into stems, optional plural markers (*-lAr*), possessive markers, and case endings (e.g., *elmalarıma* → [`elma`, `lar`, `ım`, `a`]).
- **(Noun + (Plural) + Case)**: Features 100 simpler noun forms with optional plural and case markers (e.g., *elmalara* → [`elma`, `lar`, `a`]), providing a baseline for bare nominal inflection.

Arnett and Bergen (2025) previously investigated the morphological alignment of tokenizers across typologically diverse languages to investigate whether a lack of alignment could explain performance gaps in language models. They found no evidence to suggest that morphological alignment plays a significant role. However, their operationalization of morphological alignment differs from ours as far as Turkish is concerned. While they assess whether the tokenizer separates the stem from all other suffixes combined, we investigate whether the tokenizer identifies all the different morpheme boundaries.

We evaluate the models based on their ability to approximate gold-standard morpheme boundaries for the 400 words in our dataset. To that end, we use four different metrics to quantify morphological alignment. These four metrics are: (1) the average Damerau-Levenshtein distance (Damerau, 1964) between the gold morphemes and the subwords obtained by the model, (2) the proportion of undersegmented words which refers to cases where the model produced fewer segments than the gold data, (3) the proportion of oversegmented words, and (4) the proportion of times when the model produced an output identical to the gold data.

Tokenizer	Avg. Distance	Undersegm. Items	Oversegm. Items	Exact Matches
BERTurk	2.48	82.2	0.8	5.8
EuroLLM	2.36	24.8	27.5	2.5
Goldfish	1.62	83.0	3.5	9.8
Qwen 2.5	5.50	7.0	54.2	0.5

Table 11: Morphological alignment results.

Table 11 illustrates each tokenizer's alignment to gold-standard morpheme boundaries. We observe no correlation between subword count coefficients and the latter three metrics (undersegmentation rate, oversegmentation rate, and exact match proportion). Interestingly, the ranking of models by Damerau-Levenshtein distance perfectly mirrors their ranking by subword count coefficients.

The two best-performing models and the two worst-performing models do not cluster together based on either their subword count coefficients or morphological alignment scores. This reinforces the observation that performance differences cannot be attributed to tokenizer behavior alone.