Evaluating Morphological Compositional Generalization in Large Language Models

Anonymous ACL submission

Abstract

Large language models (LLMs) have demonstrated significant progress in various natural language generation and understanding tasks. However, their linguistic generalization capabilities remain questionable, raising doubts about whether these models learn language similarly to humans. While humans exhibit compositional generalization and linguistic creativity in language use, the extent to which LLMs replicate these abilities, particularly in morphology, is under-explored. In this work, we systematically investigate the morphological generalization abilities of LLMs through the lens of compositionality. We define morphemes as compositional primitives and design a novel suite of generative and discriminative tasks to assess morphological productivity and systematicity. Focusing on agglutinative languages such as Turkish and Finnish, we evaluate several state-of-the-art instruction-finetuned multilingual models, including GPT-4 and Gemini. Our analysis shows that LLMs struggle with morphological compositional generalization particularly when applied to novel word roots, with performance declining sharply as morphological complexity increases. While models can identify individual morphological combinations better than chance, their performance lacks systematicity, leading to significant accuracy gaps compared to humans.

1 Introduction

Large language models (LLMs) have recently achieved remarkable advances in the broad domain of natural language generation and understanding tasks (Gemini, 2024; Zhao et al., 2023; Bubeck et al., 2023; Wei et al., 2022; Brown et al., 2020). However, these models have also been shown to lack strong linguistic generalization capabilities (Weissweiler et al., 2023; McCoy et al., 2023; Goldman et al., 2022; Wilson et al., 2023; Linzen, 2020; Baroni, 2019). This discrepancy casts doubt on whether language models learn a language the same way as humans do. When learning a language which is essentially a finite set of words and rules, humans exhibit linguistic creativity (Chomsky, 1965; Bergs, 2019) and compositional generalization by producing and understanding novel combinations of these grammar units which are respectively known as the *productivity* and *systematicity* aspects of compositionality (Fodor and Pylyshyn, 1988; Chomsky, 1957). While compositional generalization abilities of language models have been extensively studied (Lake and Baroni, 2018; Keysers et al., 2019; Kim and Linzen, 2020), the extent to which language models employ this ability in morphology however remains largely under-explored. Recent works evaluating morphological generalization in language models have only focused on the *productivity* aspect with a limited coverage of inflectional forms (Weissweiler et al., 2023; Anh et al., 2024).

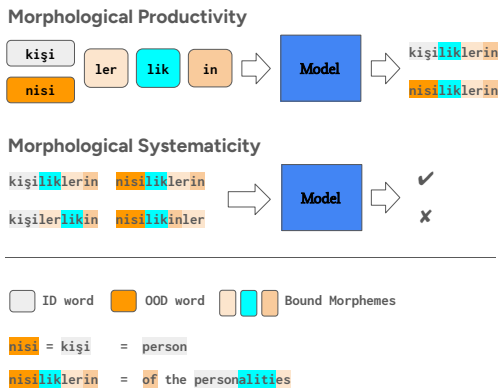

Figure 1: Our morphological generalization tasks illustrated with an example in Turkish. ID and OOD refer to in-distribution and out-of-distribution respectively. English translations are not part of the task and only shown here for illustrative purposes.

In this work, we address this gap by systematically investigating the morphological generaliza-

tion abilities of LLMs through the lens of compositionality. Following Keysers et al. (2019), we define the morphemes (smallest meaningful units in a language) [1] as the compositional primitives and design a novel suite of generative and discriminative language tasks based on the morphological combinations of these primitives. These tasks aim to test **morphological productivity** (ability to produce novel well-formed combinations of morphemes) and **systematicity** (ability to systematically understand novel combinations) respectively. Figure 1 illustrates an example of both tasks.

We evaluate several state-of-the-art instruction-finetuned large multilingual models on these tasks: GPT-4, Gemini-1.5, Aya-23 and Qwen-2.5. To ensure our findings are not language-specific, we experiment with two morphologically rich (i.e. characterized by a large number of inflectional and derivational forms) languages: Turkish and Finnish. Both languages share typological features (e.g. agglutination) despite being unrelated.

We find that LLMs lack human-like morphological compositional generalization ability in agglutinative languages despite their high performance on various tasks in these languages. Our analysis shows that morphological productivity, especially when applied to novel word roots is highly challenging for LLMs. Moreover, as the morphological complexity of words increases the model performance sharply decreases (to nearly zero) while human performance is not consistently affected. On the systematicity task, while models perform much better than chance in identifying the validity of *individual* morphological combinations, however, this behaviour is not *robust* or *systematic* i.e. models fail to *consistently* determine validity of several compositions made up of the same set of morphemes.

In summary, our contributions are as follows: **1)** We design novel morphological generalization tasks that require compositional processing. **2)** We prepare specific test suites in both Turkish and Finnish to measure morphological generalization and make these available for future research. **3)** Using our novel tasks and test suites, we conduct a systematic analysis of morphological compositional generalization abilities of LLMs. **4)** Our findings reveal a systematic gap in LLM's ability compared to humans concerning morphological generalization that also requires compositionality.

[1] In linguistics, a distinction is made between *free* morphemes that can stand alone such as words "cat" and "come" and *bound* morphemes that can only appear as part of a larger expression (e.g. "cats", "coming") such as affixes "s" and "ing".

2 Related Work

2.1 Compositional Generalization

Compositional generalization is the capacity to understand and produce novel compositions of seen primitives and is typically characterized by systematicity and productivity (Fodor and Pylyshyn, 1988; Keysers et al., 2019). Systematicity refers to the ability to understand different combinations that are made up of the same known components such as *John loves Mary* and *Mary loves John*. Productivity, on the other hand, is the ability to produce potentially infinite novel combinations of a finite number of known building blocks such as using conjunctions to construct sentences *Mary knows that John loves Mary* and *John heard that Mary knows that John loves Mary*. Past work has developed several benchmarks to measure compositional generalization abilities of neural models both in fine-tuning and in-context learning settings and has shown this task to be highly challenging (Yang et al., 2024; Lake and Baroni, 2018; Keysers et al., 2019; Kim and Linzen, 2020; An et al., 2023; Dziri et al., 2023). These benchmarks have mainly focused on synthetic sequence matching, semantic parsing, question-answering and problem-solving tasks. Our work however, investigates compositional generalization in the context of morphology.

2.2 Morphological Generalization

Morphological generalization is the ability to understand words based on their constituent parts known as *morphemes* and combine them to derive new words (Wysocki and Jenkins, 1987). Morphemes are the smallest meaningful units of language that typically correspond to word roots and affixes (i.e. prefixes, infixes and suffixes). Composing these units to construct new words can be done through *inflection* and *derivation* tasks in morphology where derivation changes the syntactic category of the words and inflection does not. These tasks have gained considerable attention as part of the SIGMORPHON's shared tasks (Cotterell et al., 2016, 2018; Vylomova et al., 2020; Kodner and Khalifa, 2022; Goldman et al., 2023) and efforts to create a universal morphology (McCarthy et al., 2020). While transformer-based models have been shown to achieve near-perfect accuracy on these

tasks (Canby et al., 2020), recent work has also found that these results are inflated due to lemma overlap pointing to a lack of generalization (Goldman et al., 2022). Other works have recently investigated the morphological capabilities of LLMs using inflection tasks and reported similarly weak performance results (Anh et al., 2024; Weissweiler et al., 2023). Similar to our study, both of these works use the popular Wug test (Berko Gleason, 1958) to evaluate the morphological generalization, however, they only focus on the productivity aspect, and their coverage of inflectional and derivational forms is limited. On the other hand, we cast the inflection and derivation tasks into the form of a compositional generalization task and evaluate models on both productivity and systematicity aspects. While focus of other works is breadth (languages from different families), we instead conduct an in-depth analysis of morphological generalization in typologically similar but unrelated languages with a large test suite covering a wide and diverse range of inflectional and derivational combinations.

3 Methodology

3.1 Tasks

We design two novel and simple compositional tasks to test morphological abilities of models. First, a **morphological productivity task** which we define as a generative task where the model is given a word root, a list of affixes (not necessarily in the correct order) and is asked to derive a meaningful word by composing the root with the affixes in the correct order. Second, a **morphological systematicity task** which we define as a binary discriminative task where the model is again given a word root, a list of affixes and a word derived from the root using the given affixes (not necessarily a meaningful word) and is asked to determine the grammatical validity of the derived word. Figure 1 illustrates these tasks with an example in Turkish.

Additionally, to measure the morphological generalization capabilities of LLMs, we take inspiration from Berko's Wug test (Berko Gleason, 1958) that is typically used to probe the inflectional and derivational morphological knowledge of children and design out-of-distribution (OOD) versions of our tasks using nonce word roots. More specifically, for each in-distribution (ID) word root in our test suite, we automatically generate a nonce word (i.e. word that does not exist in the given language) and use it in both tasks as the word root in place of the original one. However, since the model has never seen these words, to make sure the model understands the meaning of this new word, we provide the model with the original word root as a definition of the novel word root. Our generation of nonce words relies on the underlying morphophonological features and the frequency of each letter in a given language to make sure these words are plausible and inflected in the same way as the original root. Further details on nonce word generation can be found in Appendix B.

3.2 Data

We focus our study on two highly agglutinative languages, Turkish and Finnish, and prepare test suites [2] specific for our tasks in these languages. We particularly choose these languages because they are characterized by a large number of morphemes and hence require a high degree of compositional generalization ability.

Turkish Turkic languages are well-known to be highly agglutinative where the word is composed of several morphemes in addition to a root. We select Turkish as a representative of this language family in our study. To prepare our test suite we use the Bilkent Turkish Writings Dataset[3] as our base corpus which contains $6,844$ creative writings of Turkish 101 and Turkish 102 courses between 2014-2018 and hence, is full of morphologically complex words. Data statistics can be found in Appendix Table 1. We preprocess this dataset to extract words and the sentences they are found in. Then we employ a morphological analyzer for Turkish (Ozturel et al., 2019) to segment these words into a root and surface-level morphemes. To create a diverse and balanced test suite, we sample ≈ 150 examples per morpheme length 1 to 7 while maximizing the number of unique roots and morphemes (in total $1,049$ samples). Finally, we automatically generate a nonce word for each word in our test suite by relying on the fact that surface realizations of morphemes in Turkish are characterized by deterministic morphophonological processes such as vowel harmony, consonant assimilation and elision. Final data statistics and examples can be found in Appendix Tables 2 and 5 respectively. Further details on data collection can be found in Appendix D.

[2] Link to the test suites will be available in camera-ready version.

[3] https://github.com/selimfirat/bilkent-turkish-writings-dataset

Finnish We first collect a ∼1,000,000 sentence subsample of the Finnish mC4 corpus (Xue et al., 2021). We then extract unique words from the text and morphologically segment them using `omorfi` (Pirinen, 2015) and UralicNLP (Hämäläinen, 2019). After excluding words that analyzers did not cover, we manually annotate the segmentations to identify prefixes, lemmas, and affixes among the segments. We then perform stratified sampling based on the number of affixes to ensure an even range of morphological complexity in our data set. Finally, we extract sentences corresponding to each analyzed word from mC4 and validate whether they make sense. In a significant portion of cases, we notice that the raw sentences are noisy; in these cases, we opt to generate synthetic sentences using ChatGPT, which we then validate to be grammatical. Final data statistics and examples can be found in Appendix Tables 6 and 7.

4 Experiments

Setup We treat the productivity task as an open-ended task in which the model is asked to derive a word from the given root and affixes and the systematicity task as a binary classification task in which the model is asked to determine whether the given derivation is grammatically correct. For the systematicity task, we generate negative examples by producing all the combinations of morphemes attached to the same root and choosing the top four compositions (two for morpheme lengths of 1 and 2)[4] that are closest to the original valid combination measured by the Levenshtein distance. We do this to ensure our incorrect combinations are challenging enough for the model as they will be deceptively close to a plausible derivation. We also experiment with other negative example selection strategies such as random selection and a heuristic selection based on the linguistic characteristics of the given language. We describe these settings in more detail and compare the results in Section 5.5. Finally, we manually verify all the generated negative examples and fix the label of false negatives.

Models We evaluate several state-of-the-art multilingual instruction-finetuned LLMs, namely, two open-weights models, **Aya-23** (Aryabumi et al., 2024) and **Qwen-2.5** (Team, 2024), and two closed-source models, **Gemini-1.5** (Gemini, 2024) and **GPT-4** (OpenAI, 2024). We evaluate all models on all languages except for Aya-23 which officially supports Turkish, but not Finnish. [5]. We also report the performance of a *random* baseline that generates a derivation with a random combination of given morphemes (productivity task) and randomly decides whether the derivation is grammatically correct, and a *majority* baseline, which selects the most frequent label (in our case *"No"*) for the systematicity task (not applicable for the productivity task). All models are evaluated using few-shot (1, 3, and 5) in-context learning and greedy decoding since our tasks are deterministic by nature. Unless otherwise specified, prompt instructions are in English, and number of shots is set to 5 for reported results. Further details on model evaluation can be found in Appendix C.

Evaluation Metrics For the productivity task, we use **Exact Match** accuracy against the correct derivations. For the systematicity task, we report an average of **Macro-F1** scores for each sample and a **Coherence** score that measures whether the model correctly and consistently identifies the validity (or invalidity) of all derivations for a given set of morphemes. We employ this stringent metric to test the robustness of model performance similar to (Storks and Chai, 2021).

Human Evaluation We evaluate human performance on both tasks using two native speakers [6] per language, who annotate 70 and 60 samples from the Turkish and Finnish test suites, respectively. To ensure our evaluation sample is a representative sample of the entire test suite, we randomly select 10 examples per morpheme length for each test distribution. Human annotators follow the same task instructions used for model prompts and were shown five examples. We report almost perfect or substantial inter-annotator agreement measured by Cohen's kappa score (Cohen, 1960) for both tasks, languages, and test distributions (Appendix Tables 3, 4). Finally, for each task metric, we report the average score of annotators as the final human score.

[4]For 1-morpheme words, we manually annotate a negative morpheme to generate one negative option.

[5]We also experimented with recent LLMs that are instruction-finetuned specifically on Finnish such as Poro-34B (Luukkonen et al., 2024) and Ahma model series that are Llama (Touvron et al., 2023) models fine-tuned on Finnish, however, we omitted them from our analysis as they failed to follow our task prompts in both English and Finnish templates.

[6]Annotators were recruited from a Turkish and Finnish researcher community and were not compensated as they volunteered

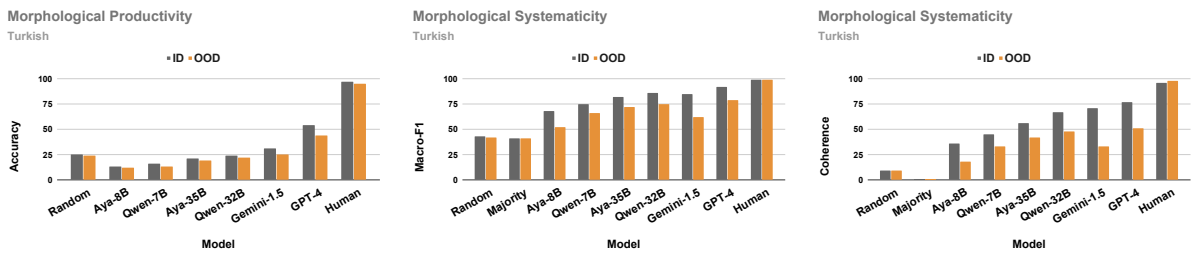

Figure 2: **Morphological productivity and systematicity task results for Turkish**. Detailed results for all shots are in Appendix Table 8.

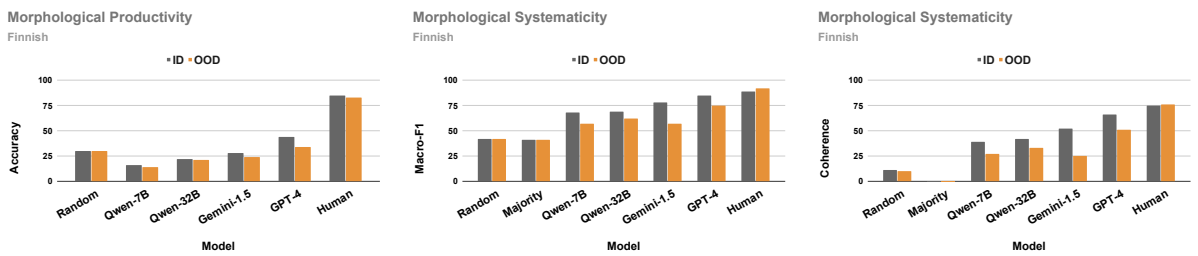

Figure 3: **Morphological productivity and systematicity task results for Finnish**. Detailed results for all shots are in Appendix Table 9.

Results Figure 2 and 3 summarize all model results for both morphological productivity and systematicity tasks evaluated respectively on the Turkish and Finnish data. We see that on the productivity task, all models except GPT-4 barely crack the random performance. While GPT-4 performs the best for both languages, it significantly lags behind the human performance (-43% and -51% in Turkish and -40.8% and -48.9% for Finnish respectively for ID and OOD data). Moreover, the GPT-4 performance gap between the ID and OOD test suites for both languages is much larger than the human gap ($\approx$ 10% vs. 3% in Turkish and 1.7% in Finnish). These results indicate that humans are much more compositionally productive in morphology and generalize more robustly to novel unseen words.

From the systematicity task results, we see that models perform much better than random and majority baselines with GPT-4 again in lead, however, the performance gap compared to humans is still significant, especially, on robustness as measured by coherence score (-19.1% and -46.5% in Turkish and -8.8% and -25.2% in Finnish respectively for ID and OOD data). The ID and OOD performance gap is also significant for all models, especially when measured by coherence score (ranging from -9.5% to -23.3% in Macro-F1 and from -13.5% to -37.2% in Coherence) while this gap is very low ($\approx 2\%$) for humans when measured by both metrics. These results show that humans are much more compositionally systematic and consistent in discriminating between correct and incorrect morphological forms made up of the same set of morphemes.

5 Analysis

5.1 Effect of Morphological Complexity

Recent works have shown that morphological complexity plays a crucial role in the morphological generalization abilities of LLMs (Anh et al., 2024). Morphological complexity is typically categorized into *integrative* (I-complexity) which refers to the predictability of inflected form and *enumerative* (E-complexity) complexity which refers to the number of cases and inflectional paradigms in language grammar (Ackerman and Malouf, 2013). While both languages we study are morphologically complex, our test suites include inflectional and derivational forms of varying length in the number of morphemes (1-7 in Turkish and 1-6 in Finnish). This allows us to study the effect of within-language E-complexity on the performance of our models. Figure 4 summarizes the GPT-4 performance for both tasks stratified by the number of bound morphemes on the Turkish data. On the productivity task, we observe a sharp downward trend (plummeting to nearly zero) in performance as the number

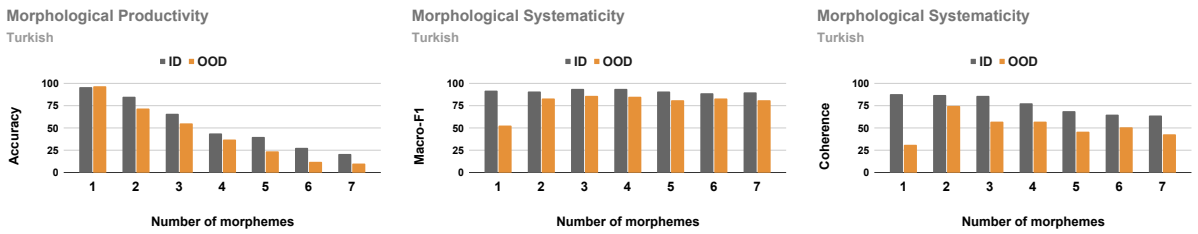

Figure 4: **GPT-4 morphological productivity and systematicity task results for Turkish stratified by number of bound morphemes**. Detailed results are in Appendix Tables 16, 17, 18. Finnish results are in Appendix Figure 11.

of morphemes increases for both ID and OOD test suites with a relatively constant gap between ID and OOD performance while humans exhibit no such dependence on complexity (Appendix Tables 16, 25). This shows that humans learn their native language robustly and can easily produce and identify long novel words while models are quite sensitive to the morphological (E-) complexity.

On the systematicity task, Macro-F1 scores for ID and OOD remain mostly unchanged as complexity increases, but coherence scores show a negative correlation with the increasing morphological complexity. We also observe a surprisingly low performance on 1-morpheme OOD words which we attribute to the varying number of negative options by morpheme length and potential shortcuts in longer morpheme words, as discussed in Appendix A.3.

5.2 Effect of Context

While our core tasks are somewhat synthetic in nature, we do also experiment with more realistic versions where we provide the model a sentence as an additional context. Specifically, we frame them as sentence completion tasks where a sentence with a blank is provided and the model is asked to fill in the blank with the correct word derived from the given word root and affixes (productivity task) or determine if the given derivation is the correct option for the blank (systematicity task).

Figure 5 summarizes the results for both productivity and systematicity tasks evaluated on the Turkish data where we provide a sentence with a blank to the model as a context (i.e. sentence completion task). This results in some improvement on the productivity task, however, we observe significant decrease in performance on the systematicity task especially for smaller models such as Aya-23 and Qwen-2.5 series and in OOD setting. This could be due to the additional complexity introduced by the extra context, however, we should note that worse performance on this task implies even stronger generalization failure since this task is more real-world and closer to the next word prediction task compared to the original context-free setup.

5.3 Effect of Tokenization

Past work has shown that suboptimal tokenizers, especially byte-pair encoding (Sennrich et al., 2016) used in GPT-4 have generally a negative effect on the morphological abilities of language models (Meyer and Buys, 2023; Bostrom and Durrett, 2020; Hofmann et al., 2021). Whether the low performance of the model on the productivity task can be attributed to the suboptimal nature of the tokenization is of interest in particular because our tasks rely on the morphologically segmented morphemes while the model utilizes byte-level tokens that are mostly English. To measure the effect of the tokenization, we ran a version of the productivity task where the morphemes provided to the model are obtained by segmenting the final derivation based on the model's own tokenizer instead of the morphologically-aligned units. Figure 6 compares the performance of the tokenizer-aligned morphemes with the morphologically-aligned morphemes on the ID test set.[7] We see that the performance in both cases is very similar to each other which points to a possibility that tokenization may not be the underlying issue behind the low performance. This finding is also consistent with some past work on exploring morphological capabilities of ChatGPT (Weissweiler et al., 2023).

5.4 Effect of Morpheme Order

Since our goal is to study the ability of LLMs to combine the morphological units in the correct order, in all of our experiments we shuffle the or-

[7] Since we use the word root as a definition for the nonce root and the tokenizer tends to break the words into meaningless chunks, we skip this experiment on the OOD test set.

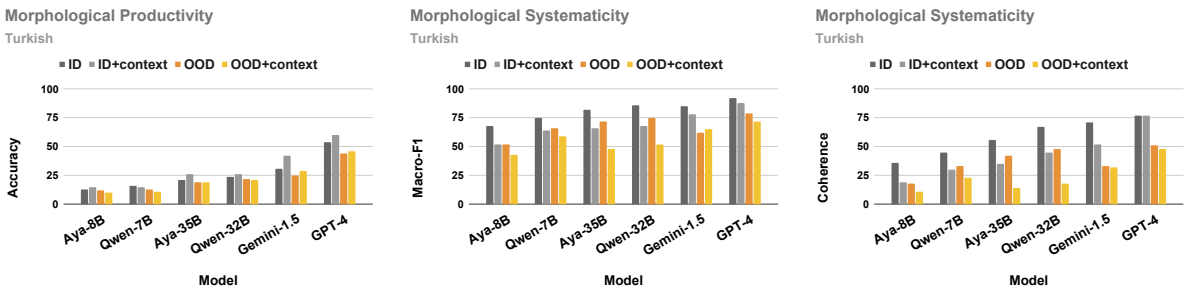

Figure 5: **Morphological productivity and systematicity task results for Turkish showing the effect of additional context**. Detailed results are in Appendix Table 36. Results for Finnish are in Appendix Figure 13.

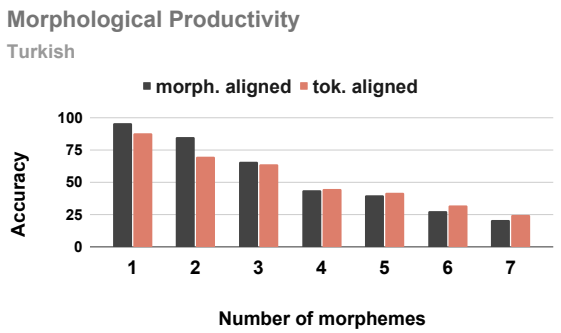

Figure 6: **GPT-4 productivity task results on the ID test suite for Turkish stratified by number of bound morphemes showing the effect of tokenization**. Detailed results are in Appendix Table 42.

der of the units in the prompts. However, given that models are sensitive to small prompt changes (Pezeshkpour and Hruschka, 2023; Zhu et al., 2024; Wang et al., 2023; Zhao et al., 2021), we also analyze the effect of changing the morpheme order on the performance of the model. To this end, we run our main experiments with all the morphemes in their correct order and report the results in Figure 7. We can see that this small change improves the performance across both tasks and models and especially, in the productivity task, the improvement can be up to 20%. This shows that models understand the tasks and can provide a correct answer by simply copying the morphemes when they are given in the correct order, however, they struggle to compose the correct order themselves. This further indicates that LLMs lack the necessary robust compositional generalization in morphology.

5.5 Effect of Negative Sample Selection

In our systematicity task, we generate negative samples (i.e. derived combinations that are not grammatically correct) by permuting the order of morphemes attached to the root. While the number of permutations is manageable for 2 or 3 morphemes (e.g., 2!=2, 3!=6), it grows rapidly with more morphemes (e.g., 6!=720). Evaluating all permutations would be ideal for robust systematicity testing, but this is infeasible due to high computational costs. Instead, we can select a subset of reasonable size to be a representative sample of all possible negative options. However, the strategy for which samples and how many to select can be somewhat arbitrary. Therefore, we experiment with three different selection strategies, and set the number of selections to four for simplicity: 1) **random** where we randomly select four negative options; 2) **language-agnostic heuristic** where we select the top four negative options that are closest to the positive option measured by Levenshtein distance (our default strategy); and 3) **language-specific heuristic** where we employ linguistic features of the tested language to filter out options that may be "too easy" for the model. We found one such heuristic for Turkish test suite based on the fact that Turkish phonology does not allow two adjacent vowels in morpheme combinations which we describe in Appendix E. We report the results of these different negative sample selection experiments in Figure 8. We see that the random selection has the highest performance on both ID and OOD test sets, followed by the language-agnostic and language-specific strategies. This implies that all our previous model results might be an upper bound and the true performance gap compared to humans is even larger than what we observe.

5.6 Error Analysis

In order to understand the limitations of language models on our tasks, we manually analyze 30 Turkish word derivations for each morpheme combination length (1-7) and for both productivity ID and

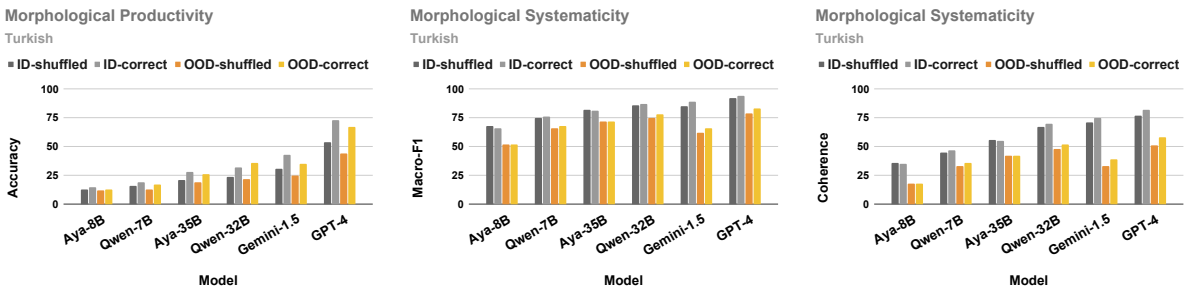

Figure 7: **Morphological productivity and systematicity task results for Turkish showing the effect of the morpheme order**. Detailed results are in Appendix Table 46. Results for Finnish are in Appendix Figure 14.

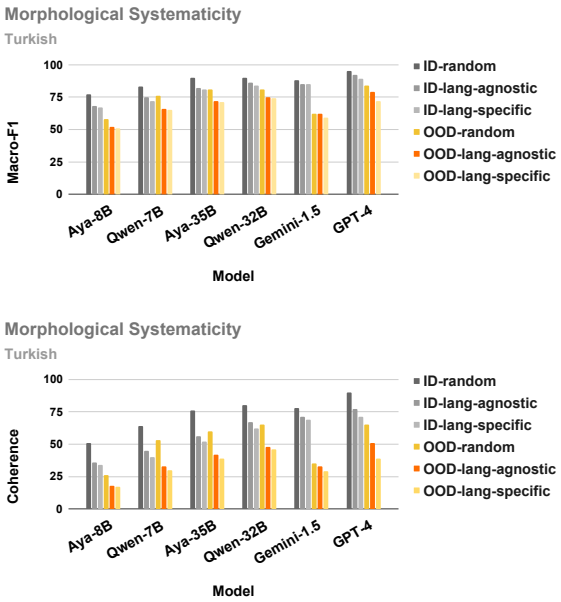

Figure 8: **Morphological systematicity task results for Turkish showing the effect of different negative sample selection strategies**. Detailed results are in Appendix Table 52.

OOD test sets resulting in a total of 178 and 185 derivations from GPT-4 that are incorrect. We annotate each generation on three criteria: 1) whether the generation is an **invalid** word (i.e. grammatically incorrect word) 2) whether the generation is **unfaithful** (i.e. generation does not follow the productivity task constraints) and 3) whether the generation includes any **hallucinations** (i.e. whether the generation has extra morphemes not mentioned in the task prompt). Our analysis shows that while on the OOD test set, GPT-4 generates a grammatically incorrect word most of the time (79%), this proportion is significantly lower for the ID test set (31%). However, on the ID test set, we observe a high unfaithfulness and hallucination ratio (91% and 67%) meaning that most of the valid generations do not follow the task constraints. On the other hand, we see lower unfaithfulness and hallucination ratios on the OOD test (75% and 52% respectively) which points to a *real word bias* also reported by (Weissweiler et al., 2023) where the model is biased toward generating frequent words for word roots existing in a given language irrespective of the underlying task. In other words, OOD setting forces the model to perform the true morphological generalization task which it fails as indicated by the higher percentage of invalid derivations. To identify the root causes of some of these errors, we analyze the GPT-4 chain-of-thought answers on the Turkish data and reveal several failure modes such as **sequential dependency errors**, **semantic misinterpretations**, **lack of grammatical knowledge**, and **unfaithful reasoning**, all of which we detail with examples in Appendix A.4. Finally, we also analyze the few errors human annotators made and find that these errors are either trivial typos or failure to notice an extra letter in a long word.

6 Conclusion

In this paper, we proposed a novel experimental paradigm to test morphological generalization abilities of large language models through compositionality. Our tasks target measuring morphological productivity and systematicity in a given language. We applied these tasks on the morphologically complex languages of Turkish and Finnish and evaluated morphological compositional generalization abilities of several state-of-the-art large language models. Our experimental results and analysis reveal a significant gap in the performance of LLMs compared to humans with respect to generalization in morphology.

Limitations

While our novel tasks are language, dataset, and model-independent, our study only focused on two agglutinative languages and a few large language models. Therefore, the applicability of our findings in other languages and models should be further studied. We also mainly focused on the grammatical validity of the words, whereas it would be equally interesting to study the capacity of LLMs to produce and understand novel semantically and pragmatically valid derivations. While we have also optimized our prompts to be as simple and maximally instructive and tested in multiple languages and in chain-of-thought setting, whether a different set of prompts would produce the same results is not clear. Finally, we only evaluate models using greedy decoding due to the deterministic nature of our tasks, however, the effect of different decoding strategies need to be explored.

References

Farrell Ackerman and Robert Malouf. 2013. Morphological organization: The low conditional entropy conjecture. *Language*, 89:429–464.

Shengnan An, Zeqi Lin, Qiang Fu, Bei Chen, Nanning Zheng, Jian-Guang Lou, and Dongmei Zhang. 2023. How do in-context examples affect compositional generalization? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11027–11052, Toronto, Canada. Association for Computational Linguistics.

Dang Anh, Limor Raviv, and Lukas Galke. 2024. Morphology matters: Probing the cross-linguistic morphological generalization abilities of large language models through a wug test. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 177–188, Bangkok, Thailand. Association for Computational Linguistics.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Marco Baroni. 2019. Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B*, 375.

Alexander Bergs. 2019. What, if anything, is linguistic creativity? *Gestalt Theory*, 41:173–183.

Jean Berko Gleason. 1958. The child's learning of english morphology. *Word*, 14.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624, Online. Association for Computational Linguistics.

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. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *Preprint*, arXiv:2303.12712.

Marc E. Canby, Aidana Karipbayeva, Bryan Lunt, Sahand Mozaffari, Charlotte Yoder, and J. Hockenmaier. 2020. University of illinois submission to the sigmorphon 2020 shared task 0: Typologically diverse morphological inflection. In *Special Interest Group on Computational Morphology and Phonology Workshop*.

N. Chomsky. 1957. *Syntactic Structures*. Janua linguarum (Mouton, Paris).: Series Minor. Mouton.

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

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The sigmorphon 2016 shared task—morphological reinflection. In *Special Interest Group on Computational Morphology and Phonology Workshop*.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jian, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaïd Harchaoui, and Yejin Choi. 2023. Faith and fate: Limits of transformers on compositionality. *ArXiv*, abs/2305.18654.

Jerry A. Fodor and Zenon W. Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1):3–71.

Team Gemini. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Omer Goldman, Khuyagbaatar Batsuren, Salam Khalifa, Aryaman Arora, Garrett Nicolai, Reut Tsarfaty, and Ekaterina Vylomova. 2023. Sigmorphon–unimorph 2023 shared task 0: Typologically diverse morphological inflection. In *Special Interest Group on Computational Morphology and Phonology Workshop*.

Omer Goldman, David Guriel, and Reut Tsarfaty. 2022. (un)solving morphological inflection: Lemma overlap artificially inflates models' performance. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 864–870, Dublin, Ireland. Association for Computational Linguistics.

Mika Hämäläinen. 2019. Uralicnlp: An nlp library for uralic languages. *Journal of Open Source Software*, 4(37):1345.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2021. Superbizarre is not superb: Derivational morphology improves BERT's interpretation of complex words. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3594–3608, Online. Association for Computational Linguistics.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. 2019. Measuring compositional generalization: A comprehensive method on realistic data. *ArXiv*, abs/1912.09713.

Najoung Kim and Tal Linzen. 2020. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105, Online. Association for Computational Linguistics.

Jordan Kodner and Salam Khalifa. 2022. Sigmorphon–unimorph 2022 shared task 0: Modeling inflection in language acquisition. *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*.

Brenden M. Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. *Preprint*, arXiv:1711.00350.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Annual Meeting of the Association for Computational Linguistics*.

Risto Luukkonen, Jonathan Burdge, Elaine Zosa, Aarne Talman, Ville Komulainen, Väinö Hatanpää, Peter Sarlin, and Sampo Pyysalo. 2024. Poro 34b and the blessing of multilinguality. *Preprint*, arXiv:2404.01856.

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Erntreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020. Unimorph 3.0: Universal morphology. In *International Conference on Language Resources and Evaluation*.

R. Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. 2023. How much do language models copy from their training data? evaluating linguistic novelty in text generation using RAVEN. *Transactions of the Association for Computational Linguistics*, 11:652–670.

Francois Meyer and Jan Buys. 2023. Subword segmental machine translation: Unifying segmentation and target sentence generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2795–2809, Toronto, Canada. Association for Computational Linguistics.

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

Adnan Ozturel, Tolga Kayadelen, and Demirsahin I. 2019. A syntactically expressive morphological analyzer for turkish. In *Proceedings of the 14th International Conference on Finite-State Methods and Natural Language Processing*, pages 65–75, Dresden, Germany. Association for Computational Linguistics.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions. *Preprint*, arXiv:2308.11483.

Tommi A Pirinen. 2015. Development and use of computational morphology of finnish in the open source and open science era: Notes on experiences with omorfi development. *SKY Journal of Linguistics*, 28:381–393.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Shane Storks and Joyce Chai. 2021. Beyond the tip of the iceberg: Assessing coherence of text classifiers. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3169–3177, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, Ran Zmigrod, Josef Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrew Krizhanovsky, Tiago Pimentel, Lucas Torroba Hennigen, Christo Kirov, Garrett Nicolai, Adina Williams, Antonios Anastasopoulos, Hilaria Cruz, Eleanor Chodroff, Ryan Cotterell, Miikka Silfverberg, and Mans Hulden. 2020. SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–39, Online. Association for Computational Linguistics.

Jiongxiao Wang, Zichen Liu, Keun Hee Park, Zhuojun Jiang, Zhaoheng Zheng, Zhuofeng Wu, Muhao Chen, and Chaowei Xiao. 2023. Adversarial demonstration attacks on large language models. *Preprint*, arXiv:2305.14950.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Preprint*, arXiv:2206.07682.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Leonie Weissweiler, Valentin Hofmann, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle, Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, Haofei Yu, Hinrich Schuetze, Kemal Oflazer, and David Mortensen. 2023. Counting the bugs in ChatGPT's wugs: A multilingual investigation into the morphological capabilities of a large language model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6508–6524, Singapore. Association for Computational Linguistics.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Michael Wilson, Jackson Petty, and Robert Frank. 2023. How abstract is linguistic generalization in large language models? experiments with argument structure. *Transactions of the Association for Computational Linguistics*, 11:1377–1395.

Katherine Wysocki and Joseph R. Jenkins. 1987. Deriving word meanings through morphological generalization. *Reading Research Quarterly*, 22(1):66–81.

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. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Haoran Yang, Hongyuan Lu, Wai Lam, and Deng Cai. 2024. Exploring compositional generalization of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 16–24.

Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Winata, and Alham Aji. 2023. Multilingual large language models are not (yet) code-switchers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12567–12582, Singapore. Association for Computational Linguistics.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. *Preprint*, arXiv:2102.09690.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *Preprint*, arXiv:2303.18223.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue

Zhang, Neil Zhenqiang Gong, and Xing Xie. 2024. Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts. *Preprint*, arXiv:2306.04528.

A Additional Analysis

A.1 Effect of Instruction Language

Since most LLMs are pre-trained on significantly more instruction data in English than other languages, we base most of our results on experiments where we use English as the prompt instruction language. However, as our data is in a different language, this results in a code-switched language which has been shown to be a challenge for large language models (Zhang et al., 2023). To measure the effect of the instruction language on the morphological generalization tasks, we run our experiments with Turkish and Finnish as the instruction language and report results for both tasks in Figure 9. We mostly observe a drop or no change in performance when the instruction language is other than English.

A.2 Effect of Chain-of-thought Reasoning

Chain-of-thought prompting has been shown to be effective in eliciting strong reasoning capabilities from LLMs (Wei et al., 2023). In order to measure the effect of this reasoning technique on LLMs' performance on our tasks, we evaluate GPT-4 (the best performing model) on both productivity and systematicity tasks in zero-shot and 5-shot chain-of-thought settings. We report the results of these experiments compared with the 5-shot standard prompting in Figure 10. We observe that while 5-shot chain-of-thought performance is better than the zero-shot chain-of-thought, it is slightly worse than or similar to the 5-shot standard prompting. To identify the causes of these errors, we manually analyze the several chain-of-thought answers which we describe in Appendix A.4.

A.3 Further details on the effect of morphological complexity

In Figure 4, we observe a surprisingly low performance ($\approx 40\%$ drop from ID performance) on the 1-morpheme OOD words, but we attribute this behaviour to the varying number of negative options available for each morpheme length and possible presence of shortcuts in larger morpheme words. We should note that we have different number of total options to discriminate for a given sample depending on the number of morphemes (for 1 and 2 morphemes, we have 2 options, for 3-7, we have 5 options). Hence, a single mistake is penalized more in the former case than in the latter. However, within the former category, we see a

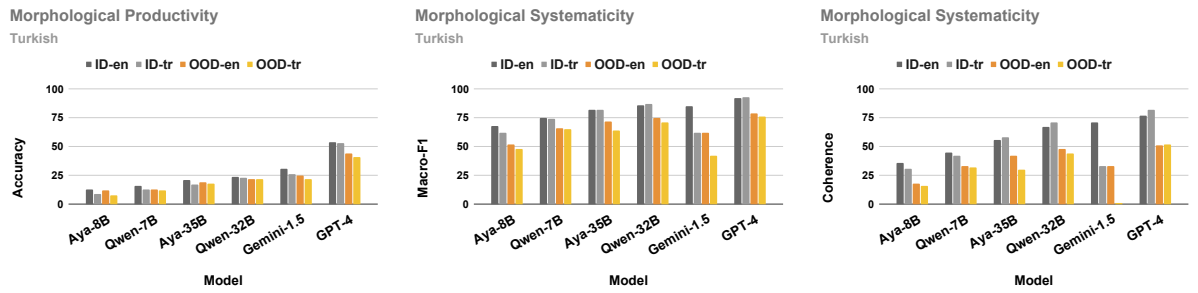

Figure 9: **Morphological productivity and systematicity task results for Turkish showing the effect of the instruction language**. Detailed results are in Table 30. Results for Finnish can be found in Figure 12.

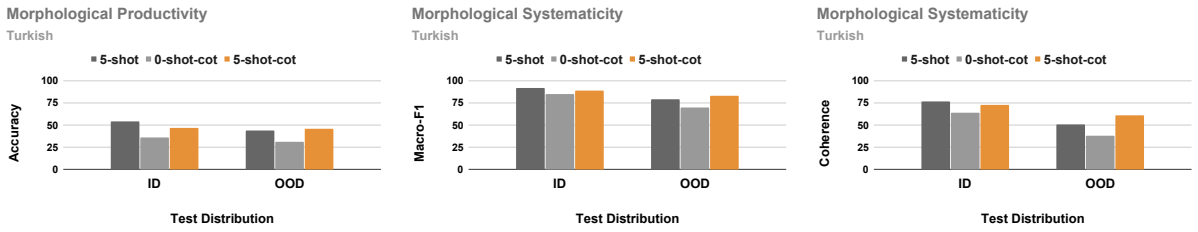

Figure 10: **GPT-4 morphological productivity and systematicity task results for Turkish showing the effect of chain-of-thought reasoning.** Detailed results are in Table 43.

much higher performance for 2-morpheme examples which might seem surprising, however, we hypothesize that this could be due to the presence of potential shortcuts for the model to exploit in the 2-morpheme case. Indeed, if we analyze the proportion of errors in both cases, we find that in the 1-morpheme case, a significant portion of errors (64%) is false negative i.e. the model identifies a nonce root with a valid morpheme as grammatically incorrect, while this is not the case in the 2-morpheme case. However, in the 2-morpheme case, the model might be exploiting the correct order of morphemes as sole evidence for the validity of the derivation while in the 1-morpheme case, there is no such shortcut and the model should understand the applicability of the given morpheme to the given word root.

A.4 Chain-of-thought Error Analysis

We randomly sample 10 examples from the 5-shot chain-of-thought experiments on the Turkish evaluation data (per morpheme length and test distribution) where GPT-4 made an error and manually analyze its answers across both tasks. Our analysis reveals the following primary types of errors:

1. **Sequential Dependency Errors**

   One common error we observe in the productivity task is due to the sequential processing of the given affixes by GPT-4. It typically starts applying the given affixes in the order they are given, however, since the affixes are typically given in shuffled order, this often results in an invalid word early on. The model, however, does not seem to realize its mistake and continues with the generation often confidently assigning meaning to the intermediate erroneous words. For example, given the word root "hedef" and affixes "-in", "-diğ" and "-le", it considers the affixes sequentially in this order by first producing "hedefin" which is valid, then "hedefindiğ" which is invalid, however, it interprets the generation as "which is the target" and finally produces "hedefindiğle" which it interprets as "with what is the target".

2. **Semantic Misinterpretations**

   Another set of errors stems from GPT-4 misinterpreting the meaning of the individual morphemes or the whole derivation. For instance, in one example, where the given morphemes are "bağır" ("to shout"), "-sa" and "-k" and GPT-4 is asked to determine the validity of the combination "bağırsak", it misinterprets this derivation as meaning "intestine" (which is also written as "bağırsak") and argues that this derivation can not be made up of the given affixes. While this reasoning is correct, the

model misses the other plausible meaning of this derivation ("as if we shout") that can be derived from the given morphemes. In another example, where the model is given the morphemes "oyna" and "-sana" and asked to produce a valid word, it misinterprets the meaning of the morpheme "-sana" as "to you" and argues that it can not be applied to the root "oyna" whereas "-sana" is a valid suffix added to verbs.

3. **Lack of Grammatical Knowledge**

   Another common pattern we see can be attributed to the lack of proper grammatical knowledge. In one example, the model is given the morphemes "uyum", "-suz", "-luk" and "-ta" and asked to determine the validity of the derivation "uyumluktasuz" which is invalid, however, the model assesses the validity of each morpheme and concludes that the combination should also be valid. In another example, it tries to add a verb suffix to a noun ("yargıyoruz"). Yet in other examples, it argues that valid affixes do not exist in the language or a valid morphological combination is not possible.

4. **Unfaithful Reasoning**

   Finally, we also observe a large set of reasoning errors due to inconsistent reasoning chains, hallucinations or unfaithful instruction following. For instance, in one example, GPT-4 concatenates the morphemes "unut" and "-alı" and yet derives "unutuluyor". In another example, it auto-corrects an invalid word ("kaldırınızdığda") to a valid word "kaldırdığınızda" and argues that the original derivation is correct.

B Nonce word generation

Turkish To automatically generate novel nonce words in Turkish (out-of-distribution words that do not exist) that are inflected the same as the original word roots, we leverage the deterministic morphophonological features of Turkish. In particular, vowel harmony and consonant assimilation in Turkish completely determines which surface forms of the meta level morphemes would apply. Furthermore, these features depend only on the last vowel and the consonant. Hence, for a given word root in Turkish, we keep its last vowel and the consonant

#words	3,775,470
#unique words	348,173
#unique roots	9,576
#unique meta affixes	103
#unique affixes	372
#unique meta affix compositions	21,930
#unique affix compositions	37,853

Table 1: Statistics of BTWD dataset in Turkish. Meta affixes refer to the bound morphemes that are not surface-realized.

#samples	1,049
#unique roots	477
#unique meta affixes	96
#unique affixes	243
#unique meta affix compositions	931
#unique affix compositions	981

Table 2: Statistics of our final test suite in Turkish. Meta affixes refer to the bound morphemes that are not surface-realized.

and randomly modify the other vowels and consonants with other vowels and consonants based on the frequency of each letter in Turkish to make sure we obtain words that would be plausible in this language. For example, if the given word root is *"sanat"*, we keep the suffix *"at"* as is and modify the prefix *"san"* by randomly replacing each vowel in it with another vowel and consonant with another consonant. This makes sure that the words inflect the same and they are of the same length. However, if the word is too short (only two letters), and there is no prefix, we generate a random prefix of length three with vowels and consonants alternating (Turkish typically doesn't allow dense consonant clusters)

Finnish The Finnish nonce word generation is done similarly to the Turkish nonce word generation, where we alter only the word root. All consonants are replaced with other consonants and vowels with other vowels that conform to the rules of Finnish vowel harmony.

C Model Evaluation

We evaluate the following state-of-the-art multilingual instruction-finetuned LLMs:

- **Aya-23** (Aryabumi et al., 2024) a powerful open-weights multilingual LLM serving 23 languages including Turkish. We evaluate

Task	Test Distribution	κ
Productivity	ID	0.94
Productivity	OOD	0.91
Systematicity	ID	0.94
Systematicity	OOD	0.99

Table 3: Human inter-annotator agreement on Turkish test suite measured by Cohen's κ score. We note that since the productivity task is an open-ended generative task, the chance agreement would be close to 0, hence κ score is equal to the raw agreement.

Task	Test Distribution	κ
Productivity	ID	0.77
Productivity	OOD	0.78
Systematicity	ID	0.75
Systematicity	OOD	0.84

Table 4: Human inter-annotator agreement on Finnish test suite measured by Cohen's κ score. We note that since the productivity task is an open-ended generative task, the chance agreement would be close to 0, hence κ score is equal to the raw agreement.

both 8B and 35B sizes of this model series, but only on Turkish dataset as Aya-23 does not officially support Finnish yet.

- **Qwen-2.5** (Team, 2024) recent open-weights multilingual LLM that has shown impressive results across various benchmarks and supports over 29 languages. We evaluate both 7B and 32B sizes of this model series in both languages.

- **Gemini-1.5** (Gemini, 2024) a closed-source multilingual LLM that supports over 40 languages including Turkish and Finnish. We evaluate the `gemini-1.5-flash` version in both languages.

- **GPT-4** (OpenAI, 2024) a closed-source multilingual LLM that supports many languages including Finnish and Turkish. We evaluate the `2024-02-15-preview` version in both languages.

Models are evaluated using in-context few-shot learning where number of shots take values in {1,3,5}. We make sure each shot has the same number of morphemes as its corresponding task example. By default, all our prompt templates are in English since LLMs are quite proficient in following instructions in this language (Wendler et al., 2024), however, we also experiment with instruction templates in Turkish and Finnish which generally show worse performance (Appendix A.1). Similarly, while by default we use the standard prompting for all experiments, we also experiment with chain-of-thought prompting (Wei et al., 2023), but find very little difference in performance (Appendix A.2). Prompts for all tasks and languages can be found in Appendix F.

D Data

Turkish Since the morphological analyzer we use to process the Turkish dataset (Ozturel et al., 2019) is based on a finite state machine relying on purely syntactic rules, it produces several alternative decompositions for some words. Hence, we further apply some language-specific heuristics to automatically filter out invalid decompositions. This preprocessing still leaves some words with multiple decompositions that can only be validated using semantics, hence, as a last step, we manually verify and determine the final segmentation of a word.

E Heuristic Negative Sample Selection For Turkish

Turkish phonology does not allow two vowels to occur together and typically employs "buffer" letters such as *"y", "s"* in between these vowels, however, blindly permuting the order of Turkish morphemes inevitably results in negative samples where two vowels may occur next to each other. We hypothesized that models might easily identify these options by exploiting the "no-two-vowel" shortcut and without considering the semantic order of morphemes. To check this hypothesis, we counted the number of GPT-4 mistakes corresponding to options that both have and don't have two vowels occurring together and found that while the model makes a mistake in around 8% (in-distribution) and 16% (out-of-distribution) of all the negative options that do not have two vowels occurring together, these ratios are only 1% and 4% when we look at the negative options that have two adjacent vowels. Motivated by this discrepancy, we designed our third heuristic-based selection strategy for Turkish such that after ranking the options by their distance to the positive option, we select the top four negative options that do not have two adjacent vowels in their morpheme composition wherever possible.

ID root (OOD root)	Affixes	ID Derivations
sohbet (şakşet)	-ler or -yin	sohbetler ✓ sohbetyin
sıra (yova)	-dan, -mış	sıradanmış ✓ sıramışdan
değer (diser)	-len, -dir, -ip	değerlendirip ✓ değeriplendir değerdirlenip değeripdirlen değerlenipdir
endişe (ödlede)	-len, -dir, -me, -mek	endişelendirmemek ✓ endişelendirmekme endişemelendirmek endişelenmedirmek endişemedirlenmek
kişi (meşi)	-leş, -tir, -me, -si, -ne	kişileştirmesine ✓ kişileştirnesime kişileştirmenesi kişileşsitirmene kişileşmetirsine
hayal (rokal)	-ler, -im, -de, -ki, -ler, -i	hayallerimdekileri ✓ hayalleriimdekiler hayalilerimdekiler hayallerimdeikiler hayallerimdekiiler
sınıf (datıf)	-lan, -dır, -ıl, -ma, -lar, -ı, -nı	sınıflandırılmalarını ✓ sınıflandırıılmalarnı sınıflardırılmalanını sınıflandırılmalarnıı sınıflandırılımalarnı

Table 5: Examples from our test suite in Turkish for each morpheme length from 1 to 7. OOD derivations can be obtained by replacing the ID root with the corresponding OOD root. Correct derivations are marked with ✓.

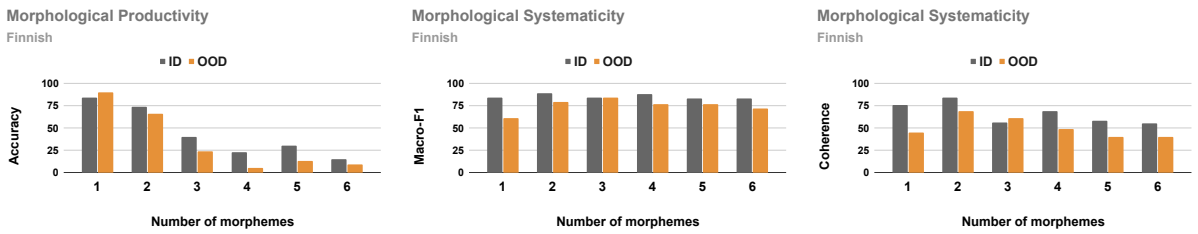

Figure 11: GPT-4 morphological productivity and systematicity task results for Finnish stratified by number of bound morphemes. Detailed results are in Tables 25, 26, 27.

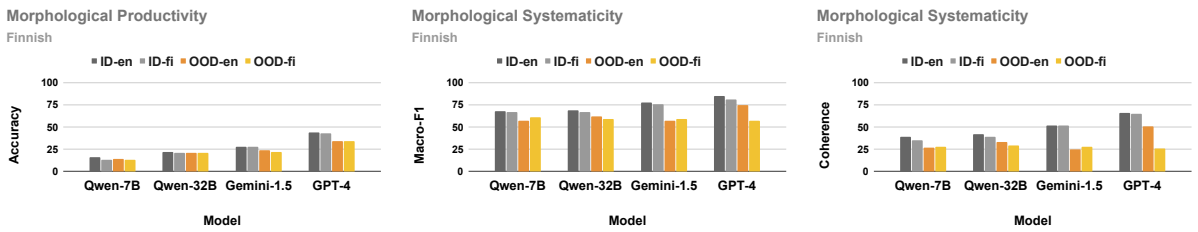

Figure 12: Morphological productivity and systematicity task results for Finnish showing the effect of the instruction language. Detailed results are in Tables 33.

#samples	480
#unique roots	406
#unique affixes	386
#unique affix compositions	365

Table 6: Statistics of our final test suite in Finnish.

F Prompts

This section lists the instruction prompts for all tasks and language templates. We present examples in one-shot setting, templates for different shots are the same with more examples. For the English language template, we provide examples in Turkish, the templates are the same for Finnish with examples in Finnish.

F.1 Templates in English

Productivity task prompt [ID root]
You are given a word root and a list of affixes (separated by comma) in Turkish and your task is to generate a grammatically correct word from this root using all the given affixes. You are allowed to use only the given affixes and each affix only once. Answer with only the generated word.

Example 1:
Word root: bulaş
Affixes: ma, sa, tır, ydı, k
Answer: bulaştırmasaydık

Example 2:
Word root: bekle
Affixes: me, di, z, n, e
Answer:

ID root (OOD root)	Affixes	ID Derivations
yöpaikka (äydainca)	-nne or -ksi	yöpaikkanne ✓ yöpaikkaksi
sano (tato)	-taan, -pas	sanotaanpas ✓ sanopastaan
petoks (seloks)	-i, -ne, -en	petoksineen ✓ petoksneien petoksneeni petoksienne petoksennei
olosuhte (olanajke)	-kuvaus, -i, -lta, -an	kuvausolosuhteiltaan ✓ kuvausolosuhteltaian kuvausolosuhteltaani kuvausolosuhteianlta kuvausolosuhteanilta
palvelu (sapsevu)	-laina, -n, -välitys, -j, -a	lainanvälityspalveluja ✓ lainanvälityspalveluaj nlainavälityspalveluja lainavälitysnpalveluja lainavälitysnpalveluaj
salaisuuks (noraekauks)	-motivaatio, -n, -nostatus, -i, -a, -ni	motivaationnostatussalaisuuksiani ✓ motivaationnostatussalaisuuksinia motivaationnostatussalaisuuksaini motivaationnostatussalaisuuksniai motivaationostatusnsalaisuuksiani

Table 7: Examples from our test suite in Finnish for each morpheme length from 1 to 6. OOD derivations can be obtained by replacing the ID root with the corresponding OOD root. Correct derivations are marked with ✓.

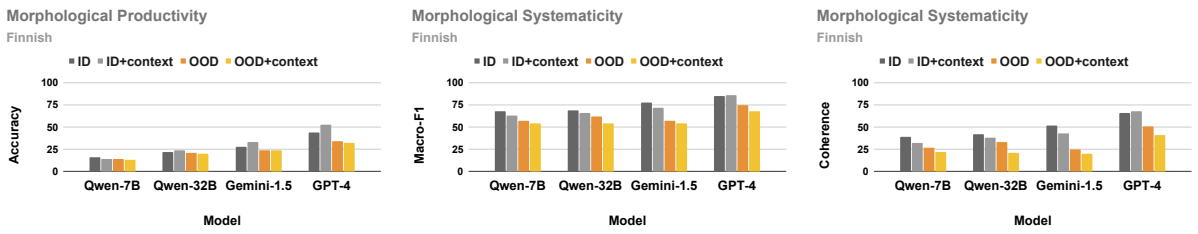

Figure 13: Morphological productivity and systematicity task results for Finnish showing the effect of additional context. Detailed results are in Table 39.

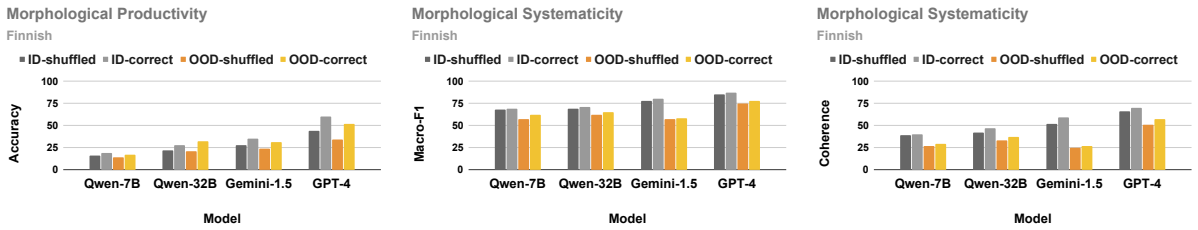

Figure 14: Morphological productivity and systematicity task results for Finnish showing the effect of the morpheme order. Detailed results are in Table 49.

Productivity task prompt [OOD root]
You are given a novel word root with its definition and a list of affixes (separated by comma) in Turkish and your task is to generate a grammatically correct word from this root using all the given affixes. You are allowed to use only the given affixes and each affix only once. Answer with only the generated word.

Example 1: Word root: lıdış
Definition: lıdış means karış in Turkish.
Affixes: sa, ydı, k, ma
Answer: lıdışmasaydık

Example 2:
Word root: ihek
Definition: ihek means emek in Turkish.
Affixes: in, imiz, ler, çi
Answer:

Systematicity task prompt [ID root]
You are given a word root, a list of affixes (separated by comma) and a word in Turkish that is derived from the given word root using the given affixes. Your task is to determine whether the derived word is grammatically correct. Answer only with Yes or No.

Example 1:
Word root: küçük
Affixes: ümüz, lüğ, den
Derived word: küçüklüğümüzden
Answer: Yes

Example 2:
Word root: evren
Affixes: sel, e, liğ
Derived word: evreneselliğ
Answer:

Models	Morph. Productivity (accuracy)		Morph. Systematicity (macro-F1)		Morph. Systematicity (coherence)	
	ID	OOD	ID	OOD	ID	OOD
majority	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0	41.2 / 41.2 / 41.2	41.2 / 41.2 / 41.2	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
random	24.6 / 25.0 / 25.0	24.7 / 24.6 / 24.2	41.8 / 41.8 / 43.5	43.0 / 42.3 / 42.3	9.1 / 9.0 / 9.4	8.5 / 9.6 / 9.0
aya-23-8b	12.8 / 13.7 / 13.3	8.8 / 11.5 / 12.3	62.0 / 64.6 / 67.5	53.9 / 49.3 / 51.5	27.9 / 31.4 / 36.0	19.1 / 15.7 / 18.4
aya-23-35b	17.4 / 19.8 / 21.0	14.6 / 17.7 / 19.3	69.9 / 80.1 / 81.8	64.6 / 71.0 / 72.1	36.8 / 52.6 / 55.8	29.2 / 39.9 / 41.8
qwen-2.5-7b	15.0 / 14.9 / 15.8	13.2 / 12.9 / 12.9	71.1 / 73.6 / 74.6	65.7 / 66.8 / 66.0	40.5 / 44.3 / 45.1	33.5 / 33.9 / 33.1
qwen-2.5-32b	22.6 / 23.7 / 24.1	21.7 / 21.8 / 21.8	77.3 / 84.7 / 85.9	53.1 / 71.3 / 75.3	56.7 / 66.3 / 66.8	18.5 / 45.7 / 48.3
gemini-1.5-flash	28.8 / 30.5 / 30.7	24.9 / 25.7 / 25.1	60.8 / 80.8 / 85.4	41.4 / 52.8 / 62.1	32.2 / 63.6 / 70.7	0.4 / 19.3 / 33.3
gpt-4	49.0 / 52.1 / 54.2	36.7 / 40.5 / 43.9	85.5 / 90.2 / 91.6	61.9 / 77.7 / 78.8	71.4 / 76.8 / 76.6	33.5 / 55.9 / 51.4
human*	97.1	95.0	98.8	99.1	95.7	97.9

Table 8: **1-shot / 3-shot / 5-shot** results for **Turkish** in **English** template for all examined models across tasks. *Due to the cost of evaluation, our human study is only evaluated on 70 randomly sampled instances per task and test distribution.

Models	Morph. Productivity (accuracy)		Morph. Systematicity (macro-F1)		Morph. Systematicity (coherence)	
	ID	OOD	ID	OOD	ID	OOD
majority	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0	40.7 / 40.7 / 40.7	40.7 / 40.7 / 40.7	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
random	29.4 / 32.3 / 29.8	32.7 / 30.2 / 30.2	42.4 / 43.3 / 42.4	42.4 / 43.7 / 42.5	10.2 / 11.7 / 10.8	10.8 / 10.8 / 9.6
qwen-2.5-7b	13.5 / 13.5 / 16.0	10.2 / 11.7 / 14.4	61.3 / 65.4 / 68.3	54.6 / 57.3 / 59.4	31.2 / 35.8 / 39.2	21.9 / 25.8 / 27.7
qwen-2.5-32b	22.5 / 21.9 / 22.3	19.2 / 19.8 / 21.3	52.0 / 65.9 / 69.0	43.6 / 54.7 / 62.2	19.0 / 39.8 / 42.5	5.2 / 22.1 / 33.1
gemini-1.5-flash	22.5 / 26.9 / 28.1	20.6 / 22.9 / 24.0	49.4 / 71.2 / 77.7	40.7 / 50.3 / 56.8	14.2 / 48.1 / 52.3	0.0 / 15.4 / 25.0
gpt-4	37.7 / 40.6 / 44.2	31.5 / 35.0 / 34.4	70.0 / 83.1 / 85.2	42.2 / 65.6 / 74.8	47.5 / 65.4 / 66.2	2.7 / 39.8 / 50.6
human*	85.0	83.3	89.4	91.7	75.0	75.8

Table 9: **1-shot / 3-shot / 5-shot** results for **Finnish** in **English** template for all examined models across tasks. *Due to the cost of evaluation, our human study is only evaluated on 60 randomly sampled instances per task and test distribution.

Systematicity task prompt [OOD root]
You are given a novel word root with its definition, a list of affixes (separated by comma) and a word in Turkish that is derived from the given word root using the given affixes. Your task is to determine whether the derived word is grammatically correct. Answer only with Yes or No.

Example 1:
Word root: eneşilvöte
Definition: eneşilvöte means üniversite in Turkish.
Affixes: niz, yse, de
Derived word: eneşilvötedeyseniz
Answer: Yes

Example 2:
Word root: yivek
Definition: yivek means yürek in Turkish.
Affixes: den, ler, iniz
Derived word: yiveklerdeniniz
Answer:

Productivity task prompt [ID root] (with context)
You are given a word root, a list of affixes (separated by comma) and a sentence with a blank (___) in Turkish and your task is to fill in the blank by generating a grammatically correct word from this root using all the given affixes. You are allowed to use only the given affixes and each affix only once. Answer with only the generated word.

Example 1:
Word root: kal
Affixes: an, lar
Sentence: giden geminin yokluğuna bir türlü inandıramaz kendilerini limanda ___
Answer: kalanlar

Example 2:
Word root: kurtar
Affixes: ecek, abil
Sentence: göç ettikten sonra diğer hemşerileri gibi mal, mülk peşinde olsa belki annesini parasızlıktan ___ belki de kızı bir fabrika köşesinde çalışmak zorunda kalmayıp daha uzun yaşayabilecekti
Answer:

Models	Number of morphemes (excl. root)						
	1	2	3	4	5	6	7
majority	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0
random	100.0 / 100.0	48.0 / 46.7	20.0 / 18.0	3.3 / 6.7	0.0 / 1.3	0.7 / 0.0	0.0 / 0.0
aya-23-8b	60.0 / 52.7	22.7 / 8.7	5.3 / 0.0	0.7 / 0.0	0.7 / 0.0	0.0 / 0.0	0.0 / 0.0
aya-23-35b	72.7 / 69.3	35.3 / 20.7	7.3 / 6.7	4.0 / 4.0	2.0 / 1.3	0.0 / 0.0	0.7 / 0.0
qwen-2.5-7b	63.3 / 64.0	26.7 / 22.0	13.3 / 6.0	0.0 / 0.7	0.0 / 0.0	0.7 / 0.0	0.7 / 0.0
qwen-2.5-32b	82.0 / 85.3	46.3 / 42.3	18.7 / 15.3	6.0 / 6.7	3.3 / 1.3	1.3 / 0.7	0.7 / 0.0
gemini-1.5-flash	86.7 / 80.0	52.7 / 44.7	36.0 / 30.7	12.0 / 10.7	7.3 / 7.3	4.7 / 0.7	2.0 / 0.0
gpt-4	95.3 / 96.7	80.7 / 65.3	62.7 / 43.3	43.8 / 31.3	27.3 / 17.6	19.3 / 0.7	13.8 / 2.1

Table 10: Morphological productivity 1-shot **ID / OOD accuracy** results for **Turkish** in **English** template for all examined models.

Models	Number of morphemes (excl. root)						
	1	2	3	4	5	6	7
majority	33.3 / 33.3	33.1 / 33.1	44.3 / 44.3	44.4 / 44.4	44.4 / 44.4	44.4 / 44.4	44.4 / 44.4
random	44.9 / 42.4	37.6 / 39.3	40.2 / 41.4	43.5 / 44.5	42.6 / 46.4	41.9 / 42.4	42.3 / 44.8
aya-23-8b	72.9 / 54.7	68.0 / 48.7	66.1 / 50.0	58.8 / 54.8	59.0 / 56.5	54.3 / 55.3	55.1 / 57.5
aya-23-35b	70.4 / 60.2	82.7 / 70.2	83.1 / 73.0	65.0 / 58.9	63.5 / 65.5	61.5 / 61.9	63.2 / 62.6
qwen-2.5-7b	71.8 / 53.1	72.9 / 64.7	73.7 / 64.4	77.7 / 74.0	73.0 / 72.3	65.5 / 66.9	63.3 / 64.4
qwen-2.5-32b	65.6 / 34.2	57.0 / 35.1	75.6 / 57.6	87.4 / 61.5	86.8 / 62.1	84.2 / 57.3	84.7 / 63.9
gemini-1.5-flash	62.4 / 33.3	58.2 / 34.0	60.1 / 44.2	59.0 / 44.8	61.9 / 44.4	56.9 / 44.3	67.2 / 44.7
gpt-4	86.7 / 36.2	69.1 / 43.8	82.5 / 61.8	88.4 / 64.2	92.2 / 78.7	88.7 / 72.0	90.6 / 76.9

Table 11: Morphological systematicity 1-shot **ID / OOD macro-F1** results for **Turkish** in **English** template for all examined models.

Systematicity task prompt [ID root] (with context) You are given a word root, a list of affixes (separated by comma), a sentence with a blank (___) and a word in Turkish that is derived from the given word root using the given affixes. Your task is to determine whether the derived word is the correct option to fill in the blank. Answer only with Yes or No.

Example 1:
Word root: küçük
Affixes: ümüz, den, lüğ
Sentence: ___ kalma bir oyuna dönüştürdük hayatımızı
Derived word: küçüklüğümüzden
Answer: Yes

Example 2:
Word root: akıl
Affixes: lan, ız, acağ
Sentence: bir şeyler yaşadıktan sonra mı ___ hep
Derived word: akılacağızlan
Answer:

Productivity task prompt [ID root] (CoT) You are given a word root and a list of affixes (separated by comma) in Turkish. Your task is to construct a grammatically correct word by appending the given affixes to the root. Use each affix exactly once. After forming a word, list each affix used in the construction of that word to verify adherence to the rules. Check the following: Ensure no affix is used more than once, confirm that all provided affixes are used, verify that no extra affixes outside the provided list are included. Think step by step and then provide your final answer within the tags <Answer>correctword</Answer>.

Example 1:
Word root: kuru
Affixes: t, muş
Answer: First, let's append the affixes to the root "kuru" in a grammatically correct order: ...<explaining the correct order of morphemes>...
Example 2:
Word root: mana
Affixes: sız, dır
Answer:

Models	Number of morphemes (excl. root)						
	1	2	3	4	5	6	7
majority	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0
random	28.0 / 25.3	18.7 / 22.0	2.0 / 1.3	4.7 / 2.0	2.7 / 2.7	4.0 / 2.7	4.0 / 3.4
aya-23-8b	62.0 / 38.0	53.3 / 26.7	20.0 / 8.7	16.7 / 12.7	18.0 / 15.3	10.7 / 13.3	14.8 / 18.8
aya-23-35b	56.7 / 45.3	74.0 / 57.3	47.3 / 29.3	19.3 / 14.0	16.0 / 20.7	20.0 / 17.3	24.2 / 20.1
qwen-2.5-7b	60.7 / 36.0	60.7 / 49.3	39.3 / 29.3	42.0 / 34.7	36.7 / 34.7	21.3 / 24.0	22.8 / 26.2
qwen-2.5-32b	49.3 / 2.7	35.6 / 2.7	54.0 / 20.7	71.3 / 28.7	66.7 / 28.7	63.3 / 19.3	56.4 / 26.8
gemini-1.5-flash	44.0 / 0.0	38.7 / 1.3	26.7 / 0.0	24.7 / 0.7	30.7 / 0.0	22.7 / 0.0	38.3 / 0.7
gpt-4	80.0 / 4.7	54.0 / 16.0	66.7 / 30.0	76.0 / 34.7	82.7 / 58.7	68.7 / 41.3	71.8 / 49.0

Table 12: Morphological systematicity 1-shot **ID / OOD coherence** results for **Turkish** in **English** template for all examined models.

Models	Number of morphemes (excl. root)						
	1	2	3	4	5	6	7
majority	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0
random	100.0 / 100.0	46.7 / 50.7	20.7 / 16.0	5.3 / 4.7	2.7 / 0.7	0.0 / 0.0	0.0 / 0.0
aya-23-8b	58.0 / 64.7	29.3 / 13.3	5.3 / 2.0	2.0 / 0.0	1.3 / 0.7	0.0 / 0.0	0.0 / 0.0
aya-23-35b	73.3 / 84.0	43.3 / 29.3	12.0 / 6.7	5.3 / 4.0	1.3 / 0.0	1.3 / 0.0	2.0 / 0.0
qwen-2.5-7b	68.7 / 61.3	24.7 / 22.7	6.7 / 5.3	2.0 / 0.7	1.3 / 0.0	0.7 / 0.0	0.0 / 0.0
qwen-2.5-32b	84.7 / 80.7	45.0 / 38.9	21.3 / 18.7	6.7 / 11.3	4.7 / 3.3	0.7 / 0.0	2.7 / 0.0
gemini-1.5-flash	84.7 / 80.0	57.3 / 50.0	37.3 / 29.3	16.7 / 9.3	10.7 / 7.3	4.0 / 2.7	2.7 / 1.3
gpt-4	94.7 / 94.7	81.3 / 68.7	64.0 / 45.2	49.3 / 34.2	30.7 / 17.6	25.3 / 11.6	19.3 / 11.7

Table 13: Morphological productivity 3-shot **ID / OOD accuracy** results for **Turkish** in **English** template for all examined models.

Productivity task prompt [OOD root] (CoT)
You are provided with a novel word root with its definition, and a list of affixes (separated by comma) in Turkish. Your task is to construct a grammatically correct word by appending the given affixes to the root. Use each affix exactly once. After forming a word, list each affix used in the construction of that word to verify adherence to the rules. Check the following: Ensure no affix is used more than once, confirm that all provided affixes are used, verify that no extra affixes outside the provided list are included. Think step by step and then provide your final answer within the tags <Answer>correctword</Answer>.

Example 1:
Word root: doru
Definition: doru means kuru in Turkish.
Affixes: t, muş
Answer: ...<explain the correct order of morphemes based on the definition>...
Example 2:
Word root: çokan
Definition: çokan means yalan in Turkish.
Affixes: la, lar
Answer:

Systematicity task prompt [ID root] (CoT)
You are given a word root, a list of affixes (separated by comma) and a word in Turkish that is derived from the given word root using the given affixes. Your task is to determine whether the derived word is grammatically correct. First, analyze how the affixes interact with the word root. Then, assess the order in which the affixes are applied and verify that this order adheres to the language's rules. Think step by step and then provide your final answer within the tags <Answer>Yes/No</Answer>.

Example 1:
Word root: kuru
Affixes: t, muş
Derived word: kurutmuş
Answer: To analyze the derived word "kurutmuş," we need to look at the affixes and how they interact with the word root "kuru."
...<explaining the correct order of morphemes>...
Example 2:
Word root: etki
Affixes: yici, le
Derived word: etkileyici
Answer:

Models	Number of morphemes (excl. root)						
	1	2	3	4	5	6	7
majority	33.3 / 33.3	33.1 / 33.1	44.3 / 44.3	44.4 / 44.4	44.4 / 44.4	44.4 / 44.4	44.4 / 44.4
random	40.0 / 45.3	40.2 / 43.1	44.4 / 41.5	40.2 / 43.4	43.9 / 40.3	43.9 / 41.8	39.8 / 41.0
aya-23-8b	75.3 / 51.8	68.0 / 43.3	64.5 / 34.2	60.3 / 44.2	64.3 / 52.7	55.6 / 55.3	64.0 / 63.3
aya-23-35b	74.9 / 57.3	83.6 / 68.2	86.2 / 78.1	79.5 / 75.3	77.3 / 77.3	78.8 / 72.8	80.3 / 68.2
qwen-2.5-7b	60.7 / 56.9	75.3 / 62.9	76.9 / 72.2	78.4 / 67.9	74.1 / 72.3	74.9 / 67.3	74.9 / 68.4
qwen-2.5-32b	76.4 / 53.8	74.0 / 60.2	87.3 / 75.6	88.6 / 78.0	91.0 / 76.6	89.0 / 76.2	86.5 / 78.7
gemini-1.5-flash	86.0 / 45.3	81.6 / 50.4	79.1 / 55.0	71.2 / 53.9	81.2 / 55.7	79.1 / 55.1	87.6 / 54.1
gpt-4	89.6 / 59.1	81.8 / 62.9	92.5 / 84.9	94.9 / 85.7	92.2 / 84.7	88.0 / 81.8	92.4 / 84.7

Table 14: Morphological systematicity 3-shot **ID / OOD macro-F1** results for **Turkish** in **English** template for all examined models.

Models	Number of morphemes (excl. root)						
	1	2	3	4	5	6	7
majority	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0
random	25.3 / 29.3	23.3 / 24.7	2.7 / 2.7	0.7 / 2.7	3.3 / 2.0	6.0 / 1.3	2.0 / 4.7
aya-23-8b	64.7 / 34.0	52.7 / 19.3	22.0 / 2.0	22.7 / 6.7	24.0 / 12.0	13.3 / 12.7	20.1 / 23.5
aya-23-35b	64.0 / 39.3	75.3 / 54.7	56.7 / 43.3	41.3 / 34.7	39.3 / 41.3	43.3 / 36.0	48.3 / 30.2
qwen-2.5-7b	44.7 / 41.3	64.7 / 46.7	39.3 / 34.0	45.3 / 26.0	36.0 / 32.7	40.0 / 28.7	40.3 / 28.2
qwen-2.5-32b	65.3 / 32.7	61.1 / 40.9	71.3 / 49.3	65.3 / 52.0	72.0 / 51.3	66.7 / 46.0	62.4 / 47.7
gemini-1.5-flash	79.3 / 20.0	72.7 / 26.0	58.0 / 18.7	45.3 / 16.0	62.0 / 18.7	56.7 / 18.0	71.1 / 17.4
gpt-4	84.7 / 40.0	72.7 / 45.3	84.0 / 65.3	84.0 / 70.0	74.7 / 61.3	63.3 / 52.7	74.5 / 56.4

Table 15: Morphological systematicity 3-shot **ID / OOD coherence** results for **Turkish** in **English** template for all examined models.

Systematicity task prompt [OOD root] (CoT)
You are given a novel word root with its definition, a list of affixes (separated by comma) and a word in Turkish that is derived from the given word root using the given affixes. Your task is to determine whether the derived word is grammatically correct. First, analyze how the affixes interact with the word root. Then, assess the order in which the affixes are applied and verify that this order adheres to the language's rules. Think step by step and then provide your final answer within the tags <Answer>Yes/No</Answer>.

Example 1:
Word root: doru
Definition: doru means kuru in Turkish.
Affixes: t, muş
Derived word: dorutmuş
Answer: ...<explain the correct order of morphemes based on the definition>...
Example 2:
Word root: imli
Definition: imli means etki in Turkish.
Affixes: yici, le
Derived word: imlileyici
Answer:

F.2 Templates in Turkish

Productivity task prompt [ID root]
Size Türkçe bir kök ve bir ek listesi (virgülle ayrılmış) verilecek ve sizden bu kökten verilen tüm ekleri kullanarak dilbilgisel olarak doğru bir kelime üretmeniz istenecek. Sadece verilen ekleri kullanabilirsiniz ve her bir ek sadece bir kez kullanılabilir. Sadece üretilen kelimeyi çıktı olarak verin.

Örnek 1:
Kök: küçük
Ekler: ümüz, lüğ, den
Cevap: küçüklüğümüzden

Örnek 2:
Kök: sevgi
Ekler: in, li, m
Cevap:

Models	Number of morphemes (excl. root)						
	1	2	3	4	5	6	7
majority	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0
random	100.0 / 100.0	44.0 / 48.7	22.7 / 15.3	7.3 / 4.7	0.7 / 0.7	0.0 / 0.0	0.0 / 0.0
aya-23-8b	60.0 / 64.0	26.7 / 19.3	3.3 / 2.7	0.7 / 0.0	1.3 / 0.0	0.7 / 0.0	0.7 / 0.0
aya-23-35b	76.0 / 87.3	46.0 / 34.7	14.7 / 10.0	5.3 / 3.3	1.3 / 0.0	2.0 / 0.0	1.3 / 0.0
qwen-2.5-7b	70.7 / 65.3	27.3 / 16.7	8.0 / 6.7	3.3 / 0.7	0.7 / 0.7	0.0 / 0.0	0.7 / 0.0
qwen-2.5-32b	83.3 / 86.7	48.3 / 39.6	20.7 / 14.7	12.0 / 10.0	2.7 / 2.0	0.0 / 0.0	2.0 / 0.0
gemini-1.5-flash	91.3 / 79.3	56.7 / 51.3	39.3 / 29.3	12.0 / 8.7	10.0 / 6.0	2.7 / 1.3	2.7 / 0.0
gpt-4	96.0 / 96.7	85.3 / 72.0	66.0 / 55.5	43.7 / 37.3	40.0 / 23.9	28.0 / 11.6	20.6 / 10.1
human	100.0 / 100.0	100.0 / 95.0	100.0 / 95.0	100.0 / 100.0	100.0 / 95.0	90.0 / 100.0	90.0 / 80.0

Table 16: Morphological productivity 5-shot **ID / OOD accuracy** results for **Turkish** in **English** template for all examined models. 1-shot and 3-shot results can be found in Tables 10, 13 respectively.

Models	Number of morphemes (excl. root)						
	1	2	3	4	5	6	7
majority	33.3 / 33.3	33.1 / 33.1	44.3 / 44.3	44.4 / 44.4	44.4 / 44.4	44.4 / 44.4	44.4 / 44.4
random	42.9 / 46.4	43.8 / 37.6	43.6 / 42.2	40.4 / 40.4	44.5 / 43.9	43.1 / 44.0	46.0 / 41.4
aya-23-8b	74.0 / 49.6	71.3 / 49.8	67.7 / 44.3	65.6 / 52.8	69.7 / 56.7	61.3 / 48.0	63.3 / 59.4
aya-23-35b	77.3 / 63.6	87.1 / 68.2	85.7 / 79.4	82.2 / 75.2	82.1 / 76.7	80.5 / 70.5	77.8 / 71.2
qwen-2.5-7b	66.0 / 58.0	76.2 / 62.7	76.2 / 71.1	78.6 / 70.4	74.4 / 68.6	76.6 / 67.1	74.5 / 63.8
qwen-2.5-32b	80.4 / 58.4	78.1 / 68.9	89.5 / 79.1	90.3 / 80.8	89.7 / 80.8	88.1 / 81.0	85.5 / 77.9
gemini-1.5-flash	86.9 / 53.3	82.9 / 55.8	83.4 / 66.3	85.5 / 60.0	84.8 / 60.5	85.5 / 66.3	89.2 / 72.3
gpt-4	92.0 / 52.7	90.9 / 82.9	94.4 / 86.1	93.9 / 84.7	90.8 / 80.6	89.1 / 82.9	90.2 / 81.4
human	100.0 / 100.0	100.0 / 96.7	100.0 / 100.0	97.2 / 100.0	100.0 / 100.0	98.8 / 97.6	95.2 / 100.0

Table 17: Morphological systematicity 5-shot **ID / OOD macro-F1** results for **Turkish** in **English** template for all examined models. 1-shot and 3-shot results can be found in Tables 11, 14 respectively.

Productivity task prompt [OOD root]
Size Türkçe yeni bir kök, onun tanımlaması ve bir ek listesi (virgülle ayrılmış) verilecek ve sizden bu kökten verilen tüm ekleri kullanarak dilbilgisel olarak doğru bir kelime üretmeniz istenecek. Sadece verilen ekleri kullanabilirsiniz ve her bir ek sadece bir kez kullanılabilir. Sadece üretilen kelimeyi çıktı olarak verin.

Örnek 1:
Kök: nıtal
Tanım: nıtal Türkçe kal anlamına gelir.
Ekler: lar, an
Cevap: nıtalanlar

Örnek 2:
Kök: rarcu
Tanım: rarcu Türkçe vurgu anlamına gelir.
Ekler: la, mış
Cevap:

Systematicity task prompt [ID root]
Size Türkçe bir kök, bir ek listesi (virgülle ayrılmış) ve bu ekleri kullanarak türetilmiş bir kelime verilecek. Sizden bu kelimenin dilbilgisel olarak doğru olup olmadığını belirlemeniz istenecek. Sadece Evet veya Hayır ile cevap verin.

Örnek 1:
Kök: küçük
Ekler: ümüz, lüğ, den
Türetilmiş kelime: küçüklüğümüzden
Cevap: Evet

Örnek 2:
Kök: sahip
Ekler: iniz, diğ, len
Türetilmiş kelime: sahipdiğinizlen
Cevap:

Models	Number of morphemes (excl. root)						
	1	2	3	4	5	6	7
majority	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0
random	24.7 / 30.0	26.0 / 20.0	2.0 / 1.3	4.7 / 1.3	2.0 / 4.7	4.0 / 4.0	2.7 / 2.0
aya-23-8b	62.7 / 32.0	58.7 / 30.7	24.7 / 7.3	30.0 / 14.7	34.0 / 15.3	18.0 / 8.0	24.2 / 20.8
aya-23-35b	68.0 / 50.7	80.7 / 53.3	56.0 / 49.3	47.3 / 36.0	49.3 / 38.7	46.0 / 30.0	43.0 / 34.9
qwen-2.5-7b	51.3 / 40.7	66.7 / 46.0	34.0 / 30.7	43.3 / 30.7	37.3 / 31.3	39.3 / 27.3	43.6 / 24.8
qwen-2.5-32b	71.3 / 40.7	67.1 / 54.4	72.0 / 48.0	72.0 / 50.0	67.3 / 53.3	62.0 / 49.3	55.7 / 42.3
gemini-1.5-flash	80.7 / 31.3	75.3 / 34.7	65.3 / 34.7	69.3 / 26.7	64.7 / 27.3	68.0 / 34.7	71.8 / 43.6
gpt-4	88.0 / 31.3	86.7 / 74.7	86.0 / 57.3	78.0 / 56.7	68.7 / 46.0	64.7 / 50.7	64.4 / 43.0
human	100.0 / 100.0	100.0 / 95.0	100.0 / 100.0	95.0 / 100.0	100.0 / 100.0	95.0 / 90.0	80.0 / 100.0

Table 18: Morphological systematicity 5-shot **ID / OOD coherence** results for **Turkish** in **English** template for all examined models. 1-shot and 3-shot results can be found in Tables 12, 15 respectively.

Models	Number of morphemes (excl. root)					
	1	2	3	4	5	6
majority	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0
random	100.0 / 100.0	43.8 / 67.5	18.8 / 10.0	7.5 / 10.0	5.0 / 5.0	1.2 / 3.8
qwen-2.5-7b	55.0 / 45.0	22.5 / 15.0	2.5 / 1.2	1.2 / 0.0	0.0 / 0.0	0.0 / 0.0
qwen-2.5-32b	72.5 / 71.2	41.2 / 32.5	18.8 / 8.8	2.5 / 2.5	0.0 / 0.0	0.0 / 0.0
gemini-1.5-flash	75.0 / 73.8	38.8 / 36.2	12.5 / 8.8	6.2 / 1.2	2.5 / 3.8	0.0 / 0.0
gpt-4	81.2 / 86.2	76.2 / 58.8	26.2 / 21.2	15.0 / 6.2	18.8 / 8.8	8.8 / 7.5

Table 19: Morphological productivity 1-shot **ID / OOD accuracy** results for **Finnish** in **English** template for all examined models.

Systematicity task prompt [OOD root]
Size Türkçe yeni bir kök, onun tanımlaması, bir ek listesi (virgülle ayrılmış) ve bu ekleri kullanarak türetilmiş bir kelime verilecek. Sizden bu kelimenin dilbilgisel olarak doğru olup olmadığını belirlemeniz istenecek. Sadece Evet veya Hayır ile cevap verin.

Örnek 1:
Kök: yivük
Tanım: yivük Türkçe küçük anlamına gelir.
Ekler: den, lüğ, ümüz
Türetilmiş kelime: yivüklüğümüzden
Cevap: Evet

Örnek 2:
Kök: minlek
Tanım: minlek Türkçe gerçek anlamına gelir.
Ekler: leş, di, me
Türetilmiş kelime: minlekleşmedi
Cevap:

Productivity task prompt [ID root] (with context)
Size Türkçe bir kök, bir ek listesi (virgülle ayrılmış) ve boşluklu (___) bir cümle verilecek ve sizden boşluğu doldurmak için bu kökten verilen tüm ekleri kullanarak dilbilgisel olarak doğru bir kelime üretmeniz istenecek. Sadece verilen ekleri kullanabilirsiniz ve her bir ek sadece bir kez kullanılabilir. Sadece üretilen kelimeyi çıktı olarak verin.

Örnek 1:
Kök: küçük
Ekler: den, ümüz, lüğ
Cümle: ___ kalma bir oyuna dönüştürdük hayatımızı
Cevap: küçüklüğümüzden

Örnek 2:
Kök: ilkokul
Ekler: da, m, ydı
Cümle: Ilk kez onun bir şiirini okuyabilme fırsatı bulduğumda, henüz daha ___ ve bu kadar farklı bir tarzla karşılaşmak beni oldukça heyecanlandırmıştı
Cevap:

Models	Number of morphemes (excl. root)					
	1	2	3	4	5	6
majority	33.3 / 33.3	33.3 / 33.3	44.4 / 44.4	44.4 / 44.4	44.4 / 44.4	44.4 / 44.4
random	41.7 / 37.1	42.1 / 46.7	42.3 / 42.7	40.7 / 41.6	42.5 / 42.2	45.0 / 44.4
qwen-2.5-7b	65.8 / 57.1	77.5 / 71.2	61.5 / 57.2	53.3 / 47.7	53.5 / 47.2	56.0 / 47.2
qwen-2.5-32b	39.2 / 34.2	58.8 / 41.7	56.6 / 49.3	55.4 / 46.5	52.2 / 45.1	49.7 / 45.1
gemini-1.5-flash	50.8 / 33.3	45.0 / 33.3	54.7 / 44.4	49.7 / 44.4	46.5 / 44.4	49.7 / 44.4
gpt-4	79.2 / 34.2	73.3 / 36.7	73.9 / 47.9	64.4 / 45.8	71.2 / 44.4	58.2 / 44.4

Table 20: Morphological systematicity 1-shot **ID / OOD macro-F1** results for **Finnish** in **English** template for all examined models.

Models	Number of morphemes (excl. root)					
	1	2	3	4	5	6
majority	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0
random	23.8 / 21.2	25.0 / 30.0	3.8 / 1.2	0.0 / 5.0	6.2 / 0.0	2.5 / 7.5
qwen-2.5-7b	50.0 / 41.2	67.5 / 57.5	26.2 / 17.5	12.5 / 5.0	15.0 / 5.0	16.2 / 5.0
qwen-2.5-32b	11.2 / 3.8	38.8 / 12.5	22.5 / 8.8	18.8 / 3.8	13.8 / 1.2	8.8 / 1.2
gemini-1.5-flash	28.7 / 0.0	17.5 / 0.0	16.2 / 0.0	10.0 / 0.0	3.8 / 0.0	8.8 / 0.0
gpt-4	68.8 / 2.5	60.0 / 5.0	52.5 / 6.2	33.8 / 2.5	45.0 / 0.0	25.0 / 0.0

Table 21: Morphological systematicity 1-shot **ID / OOD coherence** results for **Finnish** in **English** template for all examined models.

Systematicity task prompt [ID root] (with context)
Size Türkçe bir kök, bir ek listesi (virgülle ayrılmış), boşluklu (___) bir cümle ve bu ekleri kullanarak türetilmiş bir kelime verilecek. Sizden boşluğu doldurmak için bu kelimenin dilbilgisel olarak doğru olup olmadığını belirlemeniz istenecek. Sadece Evet veya Hayır ile cevap verin.

Örnek 1:
Kök: karış
Ekler: ma, sa, k, ydı
Cümle: gerçek şu ki anlayamadığımız şeylere mucize deyip ___, bugünlere belki de hiç ulaşamayacaktık
Türetilmiş kelime: karışmasaydık
Cevap: Evet

Örnek 2:
Kök: sanat
Ekler: ı, çı, lar, ndan
Cümle: tüm bu deneyimlerime ev sahipliği yapan ülke ise dünyanın en ünlü ve en çok beğenilen ___ biri olan van gogh'un doğup büyüdüğü hollanda'dan başka bir yer değil
Türetilmiş kelime: sanatçılarndanı
Cevap:

F.3 Templates in Finnish

Productivity task prompt [ID root]
Sinulle annetaan sanan sananvartalo ja luettelo pilkulla erotettuja päätteitä kielellä suomi. Tehtäväsi on luoda tästä juuresta kieliopillisesti oikea sana käyttämällä kaikkia annettuja päätteitä. Voit käyttää vain annettuja päätteitä ja kutakin päätettä vain kerran. Vastaa vain luodulla sanalla.

Esimerkki 1:
Sananvartalo: markiise
Päätteet: j, a
Vastaus: markiiseja

Esimerkki 2:
Sananvartalo: kasvattamis
Päätteet: si, ta
Vastaus:

Models	Number of morphemes (excl. root)					
	1	2	3	4	5	6
majority	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0
random	100.0 / 100.0	56.2 / 46.2	12.5 / 17.5	7.5 / 12.5	13.8 / 3.8	3.8 / 1.2
qwen-2.5-7b	58.8 / 55.0	15.0 / 11.2	5.0 / 2.5	1.2 / 0.0	1.2 / 1.2	0.0 / 0.0
qwen-2.5-32b	70.0 / 67.5	41.2 / 36.2	12.5 / 10.0	3.8 / 2.5	1.2 / 2.5	2.5 / 0.0
gemini-1.5-flash	80.0 / 76.2	51.2 / 42.5	13.8 / 7.5	5.0 / 1.2	6.2 / 6.2	5.0 / 3.8
gpt-4	81.2 / 90.0	73.8 / 62.5	32.5 / 20.0	20.0 / 13.8	26.2 / 11.2	10.0 / 12.5

Table 22: Morphological productivity 3-shot **ID / OOD accuracy** results for **Finnish** in **English** template for all examined models.

Models	Number of morphemes (excl. root)					
	1	2	3	4	5	6
majority	33.3 / 33.3	33.3 / 33.3	44.4 / 44.4	44.4 / 44.4	44.4 / 44.4	44.4 / 44.4
random	40.0 / 43.8	43.8 / 37.9	45.9 / 48.5	46.1 / 42.9	44.7 / 45.3	39.3 / 44.0
qwen-2.5-7b	62.9 / 56.7	81.2 / 77.1	67.2 / 55.4	62.8 / 52.5	57.7 / 52.0	60.4 / 50.4
qwen-2.5-32b	52.9 / 37.5	70.0 / 57.5	71.6 / 59.4	66.3 / 54.7	65.7 / 56.1	68.7 / 62.8
gemini-1.5-flash	84.2 / 50.8	70.4 / 52.1	72.1 / 55.1	64.9 / 48.8	66.8 / 46.4	69.0 / 48.6
gpt-4	81.7 / 52.5	87.9 / 65.8	84.4 / 67.7	83.0 / 68.9	78.4 / 69.7	83.1 / 69.2

Table 23: Morphological systematicity 3-shot **ID / OOD macro-F1** results for **Finnish** in **English** template for all examined models.

Productivity task prompt [OOD root]
Sinulle annetaan uusi sananvartalo, sen määritelmä sekä pilkulla eroteltu luettelo päätteitä kielellä suomi. Tehtäväsi on luoda juuresta kieliopillisesti oikea sana käyttämällä kaikkia annettuja päätteitä. Käyttä vain annettuja päätteitä ja kutakin päätettä vain kerran. Vastaa vain luodulla sanalla.

Esimerkki 1:
Sananvartalo: seloks
Määritelmä: seloks tarkoittaa petoks kielellä suomi.
Päätteet: ne, en, i
Vastaus: seloksineen

Esimerkki 2:
Sananvartalo: osivma
Määritelmä: osivma tarkoittaa ohitta kielellä suomi.
Päätteet: han, ko, a
Vastaus:

Systematicity task prompt [ID root]
Sinulle annetaan sananvartalo, pilkulla eroteltu luettelo päätteistä sekä annettuja päätteitä käyttämällä vartalosta johdettu sana kielellä suomi. Tehtäväsi on selvittää, onko johdettu sana kieliopillisesti oikein. Vastaa vain Kyllä tai Ei.

Esimerkki 1:
Sananvartalo: palauttaminen
Päätteet: n, mi, elee
Johdettu sana: mieleenpalauttaminen
Vastaus: Kyllä

Esimerkki 2:
Sananvartalo: näkyv
Päätteet: imp, in, i
Johdettu sana: näkyvimpiin
Vastaus:

Models	Number of morphemes (excl. root)					
	1	2	3	4	5	6
majority	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0
random	26.2 / 26.2	27.5 / 18.8	5.0 / 8.8	5.0 / 2.5	3.8 / 6.2	2.5 / 2.5
qwen-2.5-7b	45.0 / 41.2	72.5 / 66.2	30.0 / 15.0	27.5 / 11.2	22.5 / 12.5	17.5 / 8.8
qwen-2.5-32b	31.2 / 8.8	56.2 / 36.2	45.0 / 25.0	35.0 / 16.2	31.2 / 17.5	40.0 / 28.7
gemini-1.5-flash	76.2 / 28.7	56.2 / 28.7	43.8 / 16.2	36.2 / 7.5	37.5 / 3.8	38.8 / 7.5
gpt-4	72.5 / 32.5	82.5 / 48.8	63.7 / 38.8	61.3 / 38.8	53.8 / 41.2	58.8 / 38.8

Table 24: Morphological systematicity 3-shot **ID / OOD coherence** results for **Finnish** in **English** template for all examined models.

Models	Number of morphemes (excl. root)					
	1	2	3	4	5	6
majority	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0
random	100.0 / 100.0	50.0 / 46.2	13.8 / 17.5	7.5 / 11.2	6.2 / 2.5	1.2 / 3.8
qwen-2.5-7b	62.5 / 60.0	26.2 / 17.5	3.8 / 6.2	1.2 / 1.2	1.2 / 1.2	1.2 / 0.0
qwen-2.5-32b	72.5 / 77.5	43.8 / 33.8	10.0 / 11.2	3.8 / 2.5	1.2 / 1.2	2.5 / 1.2
gemini-1.5-flash	81.2 / 73.8	51.2 / 45.0	17.5 / 10.0	5.0 / 3.8	8.8 / 7.5	5.0 / 3.8
gpt-4	83.8 / 90.0	73.8 / 66.2	40.0 / 23.8	22.5 / 5.0	30.0 / 12.5	15.0 / 8.8
human	90.0 / 95.0	90.0 / 90.0	100.0 / 95.0	80.0 / 80.0	60.0 / 80.0	90.0 / 60.0

Table 25: Morphological productivity 5-shot **ID / OOD accuracy** results for **Finnish** in **English** template for all examined models. 1-shot and 3-shot results can be found in Tables 19, 22 respectively.

Systematicity task prompt [OOD root]
Sinulle annetaan uusi sananvartalo, sen määritelmä sekä pilkulla eroteltu luettelo päätteistä sekä uusi sana kielellä suomi, joka on johdettu annetusta sananvartalosta annettujen päätteiden avulla. Tehtäväsi on selvittää, onko johdettu sana kieliopillisesti oikein. Vastaa vain Kyllä tai Ei.

Esimerkki 1:
Sananvartalo: sätletjimsä
Määritelmä: sätletjimsä tarkoittaa järjestelmä kielellä suomi.
Päätteet: laadu, hallinta, n, n
Johdettu sana: laadunhallintasätletjimsän
Vastaus: Kyllä

Esimerkki 2:
Sananvartalo: olanajke
Määritelmä: olanajke tarkoittaa olosuhte kielellä suomi.
Päätteet: i, kuvaus, an, lta
Johdettu sana: kuvausolanajkeanilta
Vastaus:

Productivity task prompt [ID root] (with context)
Allaolevassa lauseessa (kirjoitettu kielellä suomi) on tyhjä kohta (___) joka tulee täyttää kieliopillisesti oikealla sanalla. Alla on myös sananvartalo sekä pilkulla eroteltu luettelo päätteistä. Tehtäväsi on käyttää vartaloa sekä päätteitä ja johtaa niistä kieliopillisesti oikein taivutetu sana joka sopii tyhjään kohtaan lausessaa asiayhteys/konteksti huomioonottaen. Käytä jokaista päätettä vain kerran. Vastaa vain generoidulla sanalla, älä sano mitään muuta.

Esimerkki 1:
Sananvartalo: markiise
Päätteet: a, j
Lause: ___ saatavana yksivärisinä, raidallisina ja voit myös valita haluatko markiisisi veivi- vai sähkökäyttöisenä.
Vastaus: markiiseja

Esimerkki 2:
Sananvartalo: suhteutet
Päätteet: na, tu
Lause: ___ väkilukuun, suomessa on enemmän metsää kuin missään muussa euroopan maassa.
Vastaus:

Models	Number of morphemes (excl. root)					
	1	2	3	4	5	6
majority	33.3 / 33.3	33.3 / 33.3	44.4 / 44.4	44.4 / 44.4	44.4 / 44.4	44.4 / 44.4
random	43.8 / 42.5	40.4 / 42.9	43.7 / 39.8	41.6 / 47.3	39.4 / 42.4	45.7 / 40.0
qwen-2.5-7b	65.8 / 55.4	86.7 / 75.4	65.0 / 58.1	66.2 / 54.2	62.5 / 57.7	63.5 / 55.9
qwen-2.5-32b	57.1 / 40.0	72.9 / 66.7	71.7 / 69.8	72.8 / 63.6	70.7 / 71.1	68.6 / 62.3
gemini-1.5-flash	84.2 / 49.6	80.8 / 65.4	79.6 / 62.0	75.4 / 52.8	74.9 / 53.5	71.4 / 57.7
gpt-4	84.2 / 61.3	89.2 / 78.8	84.2 / 83.8	88.0 / 76.7	83.0 / 76.5	82.7 / 71.6
human	73.3 / 86.7	93.3 / 100.0	93.2 / 97.6	95.5 / 90.9	91.2 / 90.5	89.7 / 84.4

Table 26: Morphological systematicity 5-shot **ID / OOD macro-F1** results for **Finnish** in **English** template for all examined models. 1-shot and 3-shot results can be found in Tables 20, 23 respectively.

Models	Number of morphemes (excl. root)					
	1	2	3	4	5	6
majority	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0
random	26.2 / 26.2	25.0 / 23.8	3.8 / 0.0	1.2 / 5.0	2.5 / 1.2	6.2 / 1.2
qwen-2.5-7b	48.8 / 36.2	81.2 / 65.0	28.7 / 16.2	31.2 / 15.0	22.5 / 18.8	22.5 / 15.0
qwen-2.5-32b	37.5 / 13.8	61.3 / 51.2	41.2 / 38.8	46.2 / 28.7	33.8 / 41.2	35.0 / 25.0
gemini-1.5-flash	76.2 / 28.7	71.2 / 48.8	47.5 / 26.2	43.8 / 12.5	42.5 / 12.5	32.5 / 21.2
gpt-4	76.2 / 45.0	83.8 / 68.8	56.2 / 61.3	68.8 / 48.8	57.5 / 40.0	55.0 / 40.0
human	60.0 / 80.0	90.0 / 100.0	75.0 / 90.0	85.0 / 75.0	70.0 / 60.0	70.0 / 50.0

Table 27: Morphological systematicity 5-shot **ID / OOD coherence** results for **Finnish** in **English** template for all examined models. 1-shot and 3-shot results can be found in Tables 21, 24 respectively.

Systematicity task prompt [ID root] (with context)
Allaolevassa lauseessa on tyhjä kohta (___) joka tulee täyttää kieliopillisesti oikealla sanalla. Alla on myös sananvartalo, pilkulla eroteltu luettelo päätteistä sekä niitä käyttäen annetusta vartalosta johdettu sana kielellä suomi. Tehtäväsi on päätellä, onko johdettu sana kieliopillisesti oikein, jos sen asettaa lauseen tyhjään kohtaan eli onko sana kieliopillisesti oikein taivutetu asiayhteys/konteksti huomioonottaen. Vastaa joko Kyllä tai Ei.

Esimerkki 1:
Sananvartalo: petoks
Päätteet: ne, en, i
Lause: hän paljasti koko korruptoituneen järjestelmän ___.
Johdettu sana: petoksineen
Vastaus: Kyllä
Esimerkki 2:
Sananvartalo: kannatta
Päätteet: isi, han, ko
Lause: ___ minun opiskella suomea?
Johdettu sana: kannattakoisihan
Vastaus:

Models	Morph. Productivity (accuracy)		Morph. Systematicity (macro-F1)		Morph. Systematicity (coherence)	
	ID	OOD	ID	OOD	ID	OOD
aya-23-8b	12.8 / 8.4	8.8 / 5.3	62.0 / 62.4	53.9 / 44.3	27.9 / 31.9	19.1 / 4.5
aya-23-35b	17.4 / 14.6	14.6 / 11.7	69.9 / 79.6	64.6 / 48.9	36.8 / 59.2	29.2 / 12.2
qwen-2.5-7b	15.0 / 9.3	13.2 / 10.0	71.1 / 69.9	65.7 / 59.1	40.5 / 38.1	33.5 / 23.6
qwen-2.5-32b	22.6 / 22.9	21.7 / 20.4	77.3 / 78.5	53.1 / 44.5	56.7 / 60.0	18.5 / 5.7
gemini-1.5-flash	28.8 / 21.8	24.9 / 19.3	60.8 / 45.1	41.4 / 41.2	32.2 / 5.9	0.4 / 0.0
gpt-4	49.0 / 48.5	36.7 / 38.1	85.5 / 76.3	61.9 / 50.4	71.4 / 58.6	33.5 / 15.7

Table 28: 1-shot **English / Turkish** template results for **Turkish** for all examined models across tasks.

Models	Morph. Productivity (accuracy)		Morph. Systematicity (macro-F1)		Morph. Systematicity (coherence)	
	ID	OOD	ID	OOD	ID	OOD
aya-23-8b	13.7 / 7.0	11.5 / 9.6	64.6 / 69.0	49.3 / 46.4	31.4 / 38.7	15.7 / 8.0
aya-23-35b	19.8 / 17.1	17.7 / 16.9	80.1 / 81.6	71.0 / 57.1	52.6 / 59.6	39.9 / 24.7
qwen-2.5-7b	14.9 / 11.2	12.9 / 11.4	73.6 / 73.7	66.8 / 61.4	44.3 / 44.8	33.9 / 28.2
qwen-2.5-32b	23.7 / 20.7	21.8 / 19.9	84.7 / 86.1	71.3 / 65.2	66.3 / 70.5	45.7 / 36.8
gemini-1.5-flash	30.5 / 23.0	25.7 / 20.7	80.8 / 56.9	52.8 / 41.2	63.6 / 25.3	19.3 / 0.0
gpt-4	52.1 / 51.5	40.5 / 40.6	90.2 / 88.6	77.7 / 66.3	76.8 / 76.3	55.9 / 39.9

Table 29: 3-shot **English / Turkish** template results for **Turkish** for all examined models across tasks.

Models	Morph. Productivity (accuracy)		Morph. Systematicity (macro-F1)		Morph. Systematicity (coherence)	
	ID	OOD	ID	OOD	ID	OOD
aya-23-8b	13.3 / 8.6	12.3 / 8.1	67.5 / 62.0	51.5 / 47.9	36.0 / 30.6	18.4 / 15.8
aya-23-35b	21.0 / 17.1	19.3 / 18.0	81.8 / 81.7	72.1 / 64.3	55.8 / 57.7	41.8 / 30.4
qwen-2.5-7b	15.8 / 13.0	12.9 / 12.3	74.6 / 73.5	66.0 / 65.2	45.1 / 42.4	33.1 / 32.3
qwen-2.5-32b	24.1 / 23.3	21.8 / 21.9	85.9 / 86.9	75.3 / 70.7	66.8 / 70.5	48.3 / 44.2
gemini-1.5-flash	30.7 / 25.8	25.1 / 22.1	85.4 / 61.6	62.1 / 41.6	70.7 / 32.9	33.3 / 0.7
gpt-4	54.2 / 53.1	43.9 / 40.7	91.6 / 92.7	78.8 / 76.3	76.6 / 82.1	51.4 / 51.9

Table 30: 5-shot **English / Turkish** template results for **Turkish** for all examined models across tasks. Results for 1-shot and 3-shot can be found in Tables 28 and 29.

Models	Morph. Productivity (accuracy)		Morph. Systematicity (macro-F1)		Morph. Systematicity (coherence)	
	ID	OOD	ID	OOD	ID	OOD
qwen-2.5-7b	13.5 / 7.9	10.2 / 10.8	61.3 / 56.8	54.6 / 41.2	31.2 / 23.8	21.9 / 2.1
qwen-2.5-32b	22.5 / 16.9	19.2 / 11.9	52.0 / 54.5	43.6 / 44.4	19.0 / 21.9	5.2 / 5.8
gemini-1.5-flash	22.5 / 25.2	20.6 / 21.2	49.4 / 50.2	40.7 / 40.9	14.2 / 15.6	0.0 / 0.2
gpt-4	37.7 / 37.3	31.5 / 27.7	70.0 / 59.6	42.2 / 41.6	47.5 / 30.4	2.7 / 1.2

Table 31: 1-shot **English / Finnish** template results for **Finnish** for all examined models across tasks.

Models	Morph. Productivity (accuracy)		Morph. Systematicity (macro-F1)		Morph. Systematicity (coherence)	
	ID	OOD	ID	OOD	ID	OOD
qwen-2.5-7b	13.5 / 10.4	11.7 / 10.2	65.4 / 65.0	57.3 / 55.7	35.8 / 32.5	25.8 / 22.5
qwen-2.5-32b	21.9 / 21.0	19.8 / 17.3	65.9 / 60.9	54.7 / 56.4	39.8 / 31.7	22.1 / 24.8
gemini-1.5-flash	26.9 / 24.6	22.9 / 19.2	71.2 / 74.7	50.3 / 48.0	48.1 / 51.7	15.4 / 11.9
gpt-4	40.6 / 40.8	35.0 / 32.7	83.1 / 72.0	65.6 / 45.9	65.4 / 51.0	39.8 / 8.5

Table 32: 3-shot **English / Finnish** template results for **Finnish** for all examined models across tasks.

Models	Morph. Productivity (accuracy)		Morph. Systematicity (macro-F1)		Morph. Systematicity (coherence)	
	ID	OOD	ID	OOD	ID	OOD
qwen-2.5-7b	16.0 / 13.1	14.4 / 13.3	68.3 / 66.6	59.4 / 60.1	39.2 / 35.2	27.7 / 28.1
qwen-2.5-32b	22.3 / 20.6	21.3 / 20.6	69.0 / 66.5	62.2 / 59.1	42.5 / 39.4	33.1 / 29.0
gemini-1.5-flash	28.1 / 28.1	24.0 / 21.7	77.7 / 76.2	56.8 / 59.2	52.3 / 51.9	25.0 / 28.1
gpt-4	44.2 / 42.9	34.4 / 34.4	85.2 / 81.1	74.8 / 57.2	66.2 / 65.0	50.6 / 26.2

Table 33: 5-shot **English / Finnish** template results for **Finnish** for all examined models across tasks. 1-shot and 3-shot results can be found in Tables 31 and 32.

Models	Morph. Productivity (accuracy)		Morph. Systematicity (macro-F1)		Morph. Systematicity (coherence)	
	ID	OOD	ID	OOD	ID	OOD
aya-23-8b	12.8 / 13.1	8.8 / 6.3	62.0 / 49.8	53.9 / 40.3	27.9 / 16.8	19.1 / 7.5
aya-23-35b	17.4 / 21.0	14.6 / 13.1	69.9 / 66.3	64.6 / 55.8	36.8 / 37.3	29.2 / 22.0
qwen-2.5-7b	15.0 / 13.0	13.2 / 9.8	71.1 / 61.9	65.7 / 53.6	40.5 / 29.8	33.5 / 19.7
qwen-2.5-32b	22.6 / 23.7	21.7 / 19.3	77.3 / 44.0	53.1 / 41.4	56.7 / 4.7	18.5 / 0.3
gemini-1.5-flash	28.8 / 36.1	24.9 / 24.1	60.8 / 57.4	41.4 / 43.5	32.2 / 26.2	0.4 / 3.8
gpt-4	49.0 / 59.6	36.7 / 39.9	85.5 / 71.0	61.9 / 54.5	71.4 / 49.6	33.5 / 21.4

Table 34: 1-shot **No context / With context** results for **Turkish** in **English** template for all examined models across tasks.

Models	Morph. Productivity (accuracy)		Morph. Systematicity (macro-F1)		Morph. Systematicity (coherence)	
	ID	OOD	ID	OOD	ID	OOD
aya-23-8b	13.7 / 13.6	11.5 / 9.6	64.6 / 50.3	49.3 / 42.5	31.4 / 16.8	15.7 / 9.6
aya-23-35b	19.8 / 24.3	17.7 / 17.8	80.1 / 66.3	71.0 / 51.0	52.6 / 34.9	39.9 / 16.1
qwen-2.5-7b	14.9 / 13.5	12.9 / 9.8	73.6 / 63.4	66.8 / 59.3	44.3 / 30.3	33.9 / 25.2
qwen-2.5-32b	23.7 / 24.2	21.8 / 19.3	84.7 / 61.0	71.3 / 47.3	66.3 / 33.2	45.7 / 10.2
gemini-1.5-flash	30.5 / 38.4	25.7 / 26.5	80.8 / 75.6	52.8 / 60.1	63.6 / 49.7	19.3 / 27.2
gpt-4	52.1 / 59.8	40.5 / 45.6	90.2 / 85.2	77.7 / 67.1	76.8 / 71.9	55.9 / 41.5

Table 35: 3-shot **No context / With context** results for **Turkish** in **English** template for all examined models across tasks.

Models	Morph. Productivity (accuracy)		Morph. Systematicity (macro-F1)		Morph. Systematicity (coherence)	
	ID	OOD	ID	OOD	ID	OOD
aya-23-8b	13.3 / 14.8	12.3 / 9.9	67.5 / 51.4	51.5 / 42.9	36.0 / 19.3	18.4 / 10.6
aya-23-35b	21.0 / 26.5	19.3 / 18.7	81.8 / 66.2	72.1 / 47.6	55.8 / 35.4	41.8 / 14.3
qwen-2.5-7b	15.8 / 15.0	12.9 / 11.4	74.6 / 63.9	66.0 / 58.8	45.1 / 30.3	33.1 / 23.1
qwen-2.5-32b	24.1 / 26.2	21.8 / 21.4	85.9 / 68.4	75.3 / 52.0	66.8 / 44.8	48.3 / 18.1
gemini-1.5-flash	30.7 / 41.7	25.1 / 28.7	85.4 / 77.6	62.1 / 65.3	70.7 / 51.6	33.3 / 32.2
gpt-4	54.2 / 60.0	43.9 / 46.3	91.6 / 88.4	78.8 / 72.1	76.6 / 77.2	51.4 / 48.4

Table 36: 5-shot **No context / With context** results for **Turkish** in **English** template for all examined models across tasks. 1-shot and 3-shot results can be found in Tables 34 and 35.

Models	Morph. Productivity (accuracy)		Morph. Systematicity (macro-F1)		Morph. Systematicity (coherence)	
	ID	OOD	ID	OOD	ID	OOD
qwen-2.5-7b	13.5 / 10.6	10.2 / 9.6	61.3 / 61.5	54.6 / 53.2	31.2 / 30.4	21.9 / 19.0
qwen-2.5-32b	22.5 / 22.3	19.2 / 17.3	52.0 / 43.6	43.6 / 40.7	19.0 / 4.8	5.2 / 0.0
gemini-1.5-flash	22.5 / 26.7	20.6 / 21.9	49.4 / 54.7	40.7 / 41.1	14.2 / 22.1	0.0 / 0.6
gpt-4	37.7 / 46.7	31.5 / 31.7	70.0 / 74.6	42.2 / 49.6	47.5 / 53.1	2.7 / 14.4

Table 37: 1-shot **No context / With context** results for **Finnish** in **English** template for all examined models across tasks.

Models	Morph. Productivity (accuracy)		Morph. Systematicity (macro-F1)		Morph. Systematicity (coherence)	
	ID	OOD	ID	OOD	ID	OOD
qwen-2.5-7b	13.5 / 11.0	11.7 / 10.8	65.4 / 63.5	57.3 / 55.3	35.8 / 34.0	25.8 / 21.5
qwen-2.5-32b	21.9 / 22.5	19.8 / 16.0	65.9 / 62.0	54.7 / 47.3	39.8 / 33.8	22.1 / 10.6
gemini-1.5-flash	26.9 / 32.3	22.9 / 23.3	71.2 / 72.3	50.3 / 57.3	48.1 / 43.3	15.4 / 23.3
gpt-4	40.6 / 52.3	35.0 / 33.5	83.1 / 83.8	65.6 / 65.2	65.4 / 67.9	39.8 / 39.2

Table 38: 3-shot **No context / With context** results for **Finnish** in **English** template for all examined models across tasks.

Models	Morph. Productivity (accuracy)		Morph. Systematicity (macro-F1)		Morph. Systematicity (coherence)	
	ID	OOD	ID	OOD	ID	OOD
qwen-2.5-7b	16.0 / 13.5	14.4 / 12.7	68.3 / 63.0	59.4 / 54.5	39.2 / 32.3	27.7 / 21.9
qwen-2.5-32b	22.3 / 23.8	21.3 / 20.0	69.0 / 65.5	62.2 / 54.5	42.5 / 37.9	33.1 / 20.6
gemini-1.5-flash	28.1 / 32.7	24.0 / 24.4	77.7 / 72.4	56.8 / 53.8	52.3 / 42.7	25.0 / 19.8
gpt-4	44.2 / 53.1	34.4 / 32.3	85.2 / 85.8	74.8 / 68.5	66.2 / 68.1	50.6 / 41.5

Table 39: 5-shot **No context / With context** results for **Finnish** in **English** template for all examined models across tasks. 1-shot and 3-shot results can be found in Tables 37 and 38.

Models	Number of morphemes (excl. root)						
	1	2	3	4	5	6	7
gpt-4	95.3 / 84.0	80.7 / 67.8	62.7 / 52.7	43.8 / 42.1	27.3 / 32.0	19.3 / 26.5	13.8 / 13.9

Table 40: GPT-4 morphological productivity 1-shot **morphologically aligned / tokenizer aligned accuracy** results on the ID test set for **Turkish** in **English** template.

Models	Number of morphemes (excl. root)						
	1	2	3	4	5	6	7
gpt-4	94.7 / 84.7	81.3 / 72.5	64.0 / 58.0	49.3 / 48.2	30.7 / 39.5	25.3 / 36.8	19.3 / 23.8

Table 41: GPT-4 morphological productivity 3-shot **morphologically aligned / tokenizer aligned accuracy** results on the ID test set for **Turkish** in **English** template.

Models	Number of morphemes (excl. root)						
	1	2	3	4	5	6	7
gpt-4	96.0 / 88.0	85.3 / 69.8	66.0 / 64.0	43.7 / 44.8	40.0 / 42.2	28.0 / 32.4	20.6 / 25.7

Table 42: GPT-4 morphological productivity 5-shot **morphologically aligned / tokenizer aligned accuracy** results on the ID test set for **Turkish** in **English** template. 1-shot and 3-shot results can be found in Tables 40 and 41.

Models	Morph. Productivity (accuracy)		Morph. Systematicity (macro-F1)		Morph. Systematicity (coherence)	
	ID	OOD	ID	OOD	ID	OOD
gpt-4	54.2 / 36.4 / 46.8	43.9 / 31.2 / 45.8	91.6 / 85.1 / 88.8	78.8 / 70.3 / 83.0	76.6 / 63.8 / 72.5	51.4 / 38.1 / 61.1

Table 43: GPT-4 **5-shot / 0-shot-cot / 5-shot-cot** results for **Turkish** in **English** template across tasks.

Models	Morph. Productivity (accuracy)		Morph. Systematicity (macro-F1)		Morph. Systematicity (coherence)	
	ID	OOD	ID	OOD	ID	OOD
aya-23-8b	12.8 / 14.2	8.8 / 9.4	62.0 / 62.0	53.9 / 53.5	27.9 / 28.0	19.1 / 19.4
aya-23-35b	17.4 / 22.3	14.6 / 17.9	69.9 / 69.4	64.6 / 64.3	36.8 / 36.2	29.2 / 28.2
qwen-2.5-7b	15.0 / 16.7	13.2 / 14.9	71.1 / 73.1	65.7 / 67.8	40.5 / 42.7	33.5 / 36.5
qwen-2.5-32b	22.6 / 30.0	21.7 / 29.2	77.3 / 80.4	53.1 / 58.1	56.7 / 61.0	18.5 / 25.9
gemini-1.5-flash	28.8 / 37.2	24.9 / 32.1	60.8 / 65.4	41.4 / 41.5	32.2 / 39.6	0.4 / 0.7
gpt-4	49.0 / 63.0	36.7 / 54.5	85.5 / 88.6	61.9 / 68.3	71.4 / 77.5	33.5 / 43.7

Table 44: 1-shot **Shuffled / Correct** morpheme order results for **Turkish** in **English** template for all examined models across tasks.

Models	Morph. Productivity (accuracy)		Morph. Systematicity (macro-F1)		Morph. Systematicity (coherence)	
	ID	OOD	ID	OOD	ID	OOD
aya-23-8b	13.7 / 14.9	11.5 / 12.5	64.6 / 63.8	49.3 / 49.1	31.4 / 30.7	15.7 / 16.4
aya-23-35b	19.8 / 25.7	17.7 / 23.7	80.1 / 80.2	71.0 / 73.0	52.6 / 52.2	39.9 / 41.9
qwen-2.5-7b	14.9 / 17.8	12.9 / 15.5	73.6 / 76.1	66.8 / 69.4	44.3 / 47.9	33.9 / 38.0
qwen-2.5-32b	23.7 / 30.7	21.8 / 32.1	84.7 / 86.9	71.3 / 75.7	66.3 / 70.3	45.7 / 53.0
gemini-1.5-flash	30.5 / 39.6	25.7 / 33.4	80.8 / 85.4	52.8 / 58.0	63.6 / 71.6	19.3 / 27.4
gpt-4	52.1 / 70.3	40.5 / 63.5	90.2 / 92.9	77.7 / 81.0	76.8 / 82.0	55.9 / 59.6

Table 45: 3-shot **Shuffled / Correct** morpheme order results for **Turkish** in **English** template for all examined models across tasks.

Models	Morph. Productivity (accuracy)		Morph. Systematicity (macro-F1)		Morph. Systematicity (coherence)	
	ID	OOD	ID	OOD	ID	OOD
aya-23-8b	13.3 / 15.0	12.3 / 13.0	67.5 / 66.4	51.5 / 51.8	36.0 / 34.7	18.4 / 18.1
aya-23-35b	21.0 / 28.5	19.3 / 25.8	81.8 / 81.3	72.1 / 72.3	55.8 / 55.2	41.8 / 42.3
qwen-2.5-7b	15.8 / 19.2	12.9 / 16.9	74.6 / 76.4	66.0 / 68.5	45.1 / 47.0	33.1 / 35.7
qwen-2.5-32b	24.1 / 32.3	21.8 / 36.4	85.9 / 87.5	75.3 / 78.2	66.8 / 69.6	48.3 / 52.2
gemini-1.5-flash	30.7 / 43.3	25.1 / 35.1	85.4 / 88.6	62.1 / 66.1	70.7 / 74.5	33.3 / 39.5
gpt-4	54.2 / 73.0	43.9 / 66.7	91.6 / 93.7	78.8 / 82.6	76.6 / 82.2	51.4 / 58.3

Table 46: 5-shot **Shuffled / Correct** morpheme order results for **Turkish** in **English** template for all examined models across tasks. 1-shot and 3-shot results can be found in Tables 44 and 45.

Models	Morph. Productivity (accuracy)		Morph. Systematicity (macro-F1)		Morph. Systematicity (coherence)	
	ID	OOD	ID	OOD	ID	OOD
qwen-2.5-7b	13.5 / 15.6	10.2 / 12.3	61.3 / 62.9	54.6 / 55.2	31.2 / 34.2	21.9 / 22.5
qwen-2.5-32b	22.5 / 24.8	19.2 / 23.1	52.0 / 53.1	43.6 / 44.0	19.0 / 20.4	5.2 / 5.8
gemini-1.5-flash	22.5 / 26.7	20.6 / 25.6	49.4 / 51.2	40.7 / 40.9	14.2 / 17.5	0.0 / 0.2
gpt-4	37.7 / 46.0	31.5 / 39.6	70.0 / 70.7	42.2 / 43.7	47.5 / 48.3	2.7 / 5.0

Table 47: 1-shot **Shuffled / Correct** morpheme order results for **Finnish** in **English** template for all examined models across tasks.

Models	Morph. Productivity (accuracy)		Morph. Systematicity (macro-F1)		Morph. Systematicity (coherence)	
	ID	OOD	ID	OOD	ID	OOD
qwen-2.5-7b	13.5 / 16.5	11.7 / 15.2	65.4 / 67.5	57.3 / 59.6	35.8 / 39.8	25.8 / 29.0
qwen-2.5-32b	21.9 / 25.2	19.8 / 25.0	65.9 / 66.6	54.7 / 55.4	39.8 / 41.5	22.1 / 23.1
gemini-1.5-flash	26.9 / 35.0	22.9 / 28.7	71.2 / 73.2	50.3 / 51.1	48.1 / 50.6	15.4 / 16.5
gpt-4	40.6 / 56.0	35.0 / 50.4	83.1 / 83.2	65.6 / 67.4	65.4 / 67.3	39.8 / 42.7

Table 48: 3-shot **Shuffled / Correct** morpheme order results for **Finnish** in **English** template for all examined models across tasks.

Models	Morph. Productivity (accuracy)		Morph. Systematicity (macro-F1)		Morph. Systematicity (coherence)	
	ID	OOD	ID	OOD	ID	OOD
qwen-2.5-7b	16.0 / 18.8	14.4 / 17.1	68.3 / 68.7	59.4 / 61.6	39.2 / 40.2	27.7 / 29.4
qwen-2.5-32b	22.3 / 27.9	21.3 / 31.7	69.0 / 71.3	62.2 / 64.7	42.5 / 46.7	33.1 / 37.3
gemini-1.5-flash	28.1 / 35.0	24.0 / 30.6	77.7 / 79.9	56.8 / 57.7	52.3 / 58.5	25.0 / 26.9
gpt-4	44.2 / 59.6	34.4 / 52.1	85.2 / 86.9	74.8 / 78.4	66.2 / 70.0	50.6 / 57.1

Table 49: 5-shot **Shuffled / Correct** morpheme order results for **Finnish** in **English** template for all examined models across tasks. 1-shot and 3-shot results can be found in Tables 47 and 48.

Models	Morph. Systematicity (macro-F1)		Morph. Systematicity (coherence)	
	ID	OOD	ID	OOD
aya-23-8b	74.8 / 62.0 / 60.2	59.7 / 53.9 / 51.1	46.4 / 27.9 / 25.4	26.9 / 19.1 / 16.1
aya-23-35b	83.5 / 69.9 / 67.9	74.2 / 64.6 / 61.6	59.5 / 36.8 / 33.3	43.3 / 29.2 / 26.0
qwen-2.5-7b	81.7 / 71.1 / 68.6	75.1 / 65.7 / 63.7	64.1 / 40.5 / 36.3	54.1 / 33.5 / 29.4
qwen-2.5-32b	79.7 / 77.3 / 76.6	54.0 / 53.1 / 52.8	65.5 / 56.7 / 53.8	22.0 / 18.5 / 17.7
gemini-1.5-flash	62.3 / 60.8 / 60.6	41.5 / 41.4 / 41.3	35.0 / 32.2 / 31.6	0.6 / 0.4 / 0.4
gpt-4	85.8 / 85.5 / 83.2	63.3 / 61.9 / 60.9	75.7 / 71.4 / 66.9	38.5 / 33.5 / 30.4

Table 50: 1-shot **Random / Language-agnostic / Language-specific** negative sample selection results for **Turkish** in **English** template for all examined models across tasks.

Models	Morph. Systematicity (macro-F1)		Morph. Systematicity (coherence)	
	ID	OOD	ID	OOD
aya-23-8b	74.4 / 64.6 / 61.9	53.0 / 49.3 / 47.1	45.6 / 31.4 / 28.1	21.5 / 15.7 / 13.4
aya-23-35b	88.2 / 80.1 / 78.8	80.4 / 71.0 / 71.0	72.9 / 52.6 / 48.9	60.5 / 39.9 / 38.1
qwen-2.5-7b	81.2 / 73.6 / 71.6	75.3 / 66.8 / 65.6	63.8 / 44.3 / 38.8	53.5 / 33.9 / 32.2
qwen-2.5-32b	88.3 / 84.7 / 83.4	74.5 / 71.3 / 69.8	78.3 / 66.3 / 63.6	55.4 / 45.7 / 42.4
gemini-1.5-flash	80.2 / 80.8 / 79.9	51.7 / 52.8 / 51.5	65.3 / 63.6 / 60.8	17.8 / 19.3 / 17.3
gpt-4	93.7 / 90.2 / 89.1	82.4 / 77.7 / 74.1	88.1 / 76.8 / 72.7	66.8 / 55.9 / 45.8

Table 51: 3-shot **Random / Language-agnostic / Language-specific** negative sample selection results for **Turkish** in **English** template for all examined models across tasks.

Models	Morph. Systematicity (macro-F1)		Morph. Systematicity (coherence)	
	ID	OOD	ID	OOD
aya-23-8b	77.4 / 67.5 / 66.6	58.1 / 51.5 / 50.6	51.0 / 36.0 / 33.5	25.5 / 18.4 / 17.3
aya-23-35b	89.6 / 81.8 / 80.5	80.7 / 72.1 / 70.9	75.8 / 55.8 / 51.8	60.5 / 41.8 / 38.8
qwen-2.5-7b	83.1 / 74.6 / 72.1	76.0 / 66.0 / 65.0	64.3 / 45.1 / 39.7	52.9 / 33.1 / 30.3
qwen-2.5-32b	90.1 / 85.9 / 83.9	81.3 / 75.3 / 73.6	80.3 / 66.8 / 61.8	64.7 / 48.3 / 45.8
gemini-1.5-flash	87.8 / 85.4 / 85.3	61.6 / 62.1 / 59.4	78.3 / 70.7 / 68.6	34.5 / 33.3 / 29.1
gpt-4	95.4 / 91.6 / 89.4	83.7 / 78.8 / 72.1	89.2 / 76.6 / 70.8	64.7 / 51.4 / 38.7

Table 52: 5-shot **Random / Language-agnostic / Language-specific** negative sample selection results for **Turkish** in **English** template for all examined models across tasks.