Extracting Linguistic Information from Large Language Models: Syntactic Relations and Derivational Knowledge

Tsedeniya Kinfe Temesgen[1,2] **and Marion Di Marco**[1] **and Alexander Fraser**[1,2]
[1]School of Computation, Information and Technology,
Technische Universität München (TUM)
[2]Munich Center for Machine Learning
{tsedeniya.temesgen|marion.dimarco|alexander.fraser}@tum.de

Abstract

This paper presents a study of the linguistic knowledge and generalization capabilities of Large Language Models (LLMs), focusing on their morphosyntactic competence. We design three diagnostic tasks: (i) labeling syntactic information at the sentence level - identifying subjects, objects, and indirect objects; (ii) derivational decomposition at the word level - identifying morpheme boundaries and labeling the decomposed sequence; and (iii) in-depth study of morphological decomposition in German and Amharic. We evaluate prompting strategies in GPT-4o and LLaMA 3.3-70B to extract different types of linguistic structures for typologically diverse languages. Our results show that GPT-4o consistently outperforms LLaMA in all tasks; however, both models exhibit limitations and show little evidence of abstract morphological rule learning. Importantly, we show strong evidence that the models fail to learn underlying morphological structures. Therefore, raising important doubts about their ability to generalize.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable success across a wide range of natural language understanding and generation benchmarks (OpenAI et al., 2024; Team et al., 2024; Grattafiori et al., 2024). Beyond excelling at downstream tasks, LLMs have prompted growing interest in characterizing the nature and extent of their underlying linguistic knowledge. Integrating explicit linguistic knowledge has been shown to significantly enhance model performance, particularly in low-resource languages where limited data hampers the ability to learn patterns solely from raw text. In this paper, we explore strategies to extract linguistic information from LLMs across low- and high-resource languages, focusing on core morphosyntactic tasks as a way to study their underlying syntactic and morphological representations across diverse linguistic contexts.

Recent work has investigated the linguistic competencies of LLMs across morphology, syntax, and semantics. Blevins et al. (2023) and Brown et al. (2020) found that GPT-3 exhibits substantial syntactic and semantic competence beyond mere memorization. Layer-wise analyses by He et al. (2024) on GPT-2 models show that lower layers encode morphological and syntactic features, while deeper layers capture more abstract semantic information; they also highlight challenges for LLMs at the syntax-semantics interface and in handling morphological phenomena. Findings by Waldis et al. (2024) suggest that while models perform well on formal linguistic structures (morphology and syntax), they continue to struggle with semantics, reasoning, and discourse.

Another strand of research concerns the generalization abilities with regard to generating novel forms. Studies including Anh et al. (2024), Di Marco and Fraser (2024), Weissweiler et al. (2023), and Goldman et al. (2022) have shown that while LLMs can often handle common morphological patterns, they struggle with more complex or nonce words, particularly in low-resource languages. Weller-Di Marco and Fraser (2024) argue that LLMs often rely on ad-hoc interpreting surface-level word parts rather than systematically applying structured linguistic rules. In addition, work by Ismayilzada et al. (2025) and Kodner et al. (2023) indicates that morphological generalization remains a major limitation for current models.

Linguistic competence has been shown to significantly improve tasks like machine translation (MT), especially for low-resource languages (e.g. Tanzer et al. (2024); Ramos et al. (2024); Zhang et al. (2024)). These findings emphasize that the incorporation of deep linguistic knowledge, such as morphology and syntax, can bridge major gaps in low-resource scenarios. We take this empirical

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 27210–27226
November 4-9, 2025

support as motivation to explore strategies to obtain structured linguistic knowledge from LLMs. Moreover, many studies focus primarily on English; we evaluate linguistic competencies in a diverse set of languages.

In this work, we aim to systematically evaluate LLMs' abilities in **(i) Labeling Syntactic Information** at the sentence level - identifying and labeling subcategorized phrases of a given verb - namely, subjects, objects, and indirect objects. This task probes the models' morphosyntactic competence, as it requires integrating syntactic structure with morphological cues in context. **(ii) Morphological Decomposition and Analysis** of complex derivational words, focusing on identifying morpheme boundaries and labeling each morpheme's role (e.g., stem, derivational affixes). **(iii) In-depth study of Morphological Decomposition in German and Amharic** - a variant of (ii), a detailed analysis of two typologically distinct languages. Assessing these abilities provides insight into the extent to which LLMs apply structured linguistic knowledge, as well as into the possibility to obtain structured information by means of prompting, such that it can be used for educational purposes or for the improvement of a downstream task.

We evaluate state-of-the-art instruction-tuned LLMs: GPT-4o and LLaMA 3.3-70B (referred to as LLaMA) models on 10 typologically and morphologically diverse languages. Our language set includes high-resource Italian and French (Romance), English and German (Germanic), and Polish (Slavic), as well as low-resource Latvian and Lithuanian (Baltic), Upper Sorbian (Slavic), and Amharic and Maltese (Semitic). We use a few-shot prompting setup, and include zero-shot results for comparison.

Our findings indicate that GPT-4o demonstrates stronger linguistic competence than LLaMA across all tasks. Both models achieve better performance on the syntactic labeling task as the number of examples increases. However, gains in the morphological decomposition task, evaluated across five selected languages, are inconsistent. Syntactic labeling primarily benefits from contextual cues, while the morphological decomposition task relies on the recognition of morphological structure of a word. A deeper analysis of Amharic and German reveals specific challenges. For Amharic, models struggle with the root-and-pattern morphological system, particularly where accurate segmentation requires separating vowels from consonants to identify derivational morphemes. For German, we observe structural inconsistencies in morphological tag sequences, which, however, improves when increasing the number of few-shot examples, suggesting that models learn linguistic patterns with more examples.

In summary, our work makes three contributions. (i) We design an evaluation suite to assess and extract core linguistic knowledge from LLMs for both high- and low-resource languages, and investigate few-shot prompting strategies on two state-of-the-art LLMs (GPT-4o and LLaMA). (ii) We construct new datasets for German and Amharic using morphological analyzers, for Task 3, that enable us to study complex morphological phenomena. (iii) Finally, we release all datasets[1] to support future research.

2 Related work

2.1 LLMs Linguistic Knowledge

While LLMs exhibit strong surface-level fluency, their morphological competence remains under scrutiny, particularly due to challenges introduced by subword segmentation. Ismayilzada et al. (2025) studied models' ability to produce and understand novel combinations of morphemes with a focus on instruction-tuned LLMs (GPT-4, Gemini-1.5, Aya-23, and Qwen-2.5) for Turkish and Finnish. Their results show that GPT-4 performs best among the evaluated models, yet still falls substantially short of human performance.

The authors study tokenization as one factor contributing to the models falling behind, but find that performance is related to the order of the morphemes provided in the prompts, which further indicates that LLMs lack the necessary robust compositional generalization in morphology. Anh et al. (2024) examine the generalizability of GPT-4 and GPT-3.5 across six languages. Their findings indicate models' morphological abilities are more influenced by the irregularity of a language's morphology than by its inflectional richness or the amount of training data. Similarly, Weller-Di Marco and Fraser (2024) assesses GPT-3.5's understanding of German compound words and derivational morphology. While the model succeeds at identifying the components of compound words, it struggles to recognize ill-formed derivations. Similarly,

[1] https://github.com/TsedeniyaTemesgen/Extracting-Linguistic-Information-from-LLMs.git, released under CC-BY license.

Weissweiler et al. (2023) investigates the morphological capabilities of ChatGPT by evaluating it on nonce words in English, German, Tamil, and Turkish. By using nonce words (novel word forms absent from the training data), the study aimed to test the models ability to generalize morphological rules rather than relying on memorized vocabulary. Waldis et al. (2024) present a comprehensive study on the linguistic competence of LLMs, finding that model performance is influenced by factors like architecture, model size, and instruction tuning.

2.2 Linguistic Information in NLP Tasks

Many NLP tasks (Yang, 2021; Xu et al., 2021; Bai et al., 2021; Li et al., 2021), have been shown to benefit from the inclusion of different types of linguistic information.

Machine translation (MT) in particular remains a key challenge for low-resource languages where the scarcity of parallel corpora limits the effectiveness of LLMs. Recent work emphasizes that incorporating linguistic knowledge can substantially improve MT performance under such constraints. Tanzer et al. (2024) introduce a benchmark that tests LLMs on translating between English and Kalamang using only a single grammar book, evaluated on twelve models including LLaMA variants, GPT, and Claude 2. Although the models do not reach human performance, the results improve when words are paired with morphological information or context. Ramos et al. (2024) explores ways to inject grammatical structure into MT pipelines using glosses. Their approach tested across high-, mid-, and low-resource languages, and showed that prompting with glosses improves translation accuracy with LLaMA-3 70B. Zhang et al. (2024) investigates various NLP tasks, including MT, by injecting grammar books, dictionaries, and morphologically analyzed texts for endangered languages, leading to considerable improvements on GPT-4.

These studies show that various types of linguistic input can significantly improve downstream tasks such as machine translation, underscoring the importance of structured linguistic information, and thus raising the question of whether multilingual LLMs can be leveraged to obtain such information.

2.3 Designing an Evaluation Dataset

Reliable evaluation of morphological generalization requires carefully designed datasets that minimize overlap between training and test sets. Without this control, models may appear to perform well by memorizing seen word forms rather than true generalization. Recent studies have highlighted this issue: Kodner et al. (2023) investigates lemma-feature overlap, and Goldman et al. (2022) introduces a lemma-split approach, where no lemmas are shared between training and test data. Both studies report that models struggle with unseen word forms and feature combinations, even in languages where generalization should be feasible, underscoring the need for careful dataset design. Following this line of work, we adopt a lemma-split approach in our morphological decomposition task by preparing data in which there is no lemma overlap between the few-shot examples and the test set.

3 Methodology

We evaluate two instruction-tuned LLMs: GPT-4o and LLaMA 3.3-70B via API access. To ensure deterministic outputs, we set the temperature to 0 and generate a single response per prompt using the `chat.completions.create function`. We design prompt instructions and experiments with N-shot settings (N = 0, 1, 3, 5, 10).

In the following section, we present three tasks designed to evaluate different aspects of linguistic knowledge. **Task 1 – Labeling syntactic information** assesses the models' ability to identify and label **subjects**, **objects**, and **indirect objects** of a given verb in a sentence, targeting morpho-syntactic competence. **Task 2 – Morphological decomposition** evaluates the ability to segment derivational words into constituent morphemes and assign functional labels, probing morphological knowledge. **Task 3 – Morphological decomposition in German and Amharic** extends Task 2 with a deeper, language-specific analysis to examine model performance in morphologically rich languages.

3.1 Task 1: Labeling Syntactic Information

We use data from the Universal Dependencies (UD) treebanks[2] for 10 diverse languages: Amharic (amh), English (eng), French (fra), German (deu), Italian (ita), Latvian (lav), Lithuanian (lit), Maltese (mlt), Polish (pol), and Upper Sorbian (hsb); see A.1 for datasets details.

[2]https://universaldependencies.org/

To ensure sufficient syntactic context, we filter out sentences with four or fewer words. For each remaining sentence, we extract all arguments associated with each verb: subject, object, and indirect object. We then keep only those cases where a subject is present, discarding any instances where an object or indirect object appears without one. We then retain sentences whose length is close to the average, to minimize bias arising from trivially short inputs. For datasets without predefined train, development, and test splits, we allocate 0.25/0.25/0.50 for train/dev/test (See B.3). Finally, we structured the dataset as in Paolini et al. (2021), where each word is labeled with its corresponding argument (See Appendix C).

Despite being typologically different, the datasets for all languages are created uniformly, with the exception of Amharic, for which an additional variant is introduced. Amharic exhibits a root-and-pattern morphological structure in which subject, object, and indirect object markers are expressed as bound morphemes (see Section A.2 for language properties). These morphemes are most commonly attached to verbs, encoding multiple syntactic functions within a single word (see example in B.1). Although they are often segmented and annotated as separate syntactic words in the UD dataset, they do not function as independent lexical items and typically appear alongside an overt subject.

To capture this distinction, we prepared two versions of the Amharic UD dataset. In the first variant, referred to as **Amharic**(*amh*), words are to be labeled with one or more syntactic roles. In the second one, **Amharic-morph**(*amh-morph*), words are segmented into individual morphemes, each of which is then labeled to indicate subject and/or indirect object.

Few-shot examples: We select N examples to cover all argument combinations: subject only, subject with object, subject with indirect object, and all three combined, thus exposing the model to the full range of syntactic patterns and argument structures. For example, in English, setting $N = 1$ yields four training examples, with one selected from each category. (See B.3 few-shot selection).

Prompt: We design instructions that require the model to label sentences with the correct argument roles: subject, object, and indirect object. The model is explicitly instructed not to provide explanations or to modify the sentence. In addition, models are prompted to label the rightmost word in case of multi-word phrases. (See Table 3 for prompt instruction)

3.2 Task 2: Morphological decomposition

We studied five languages (*eng, fra, deu, lav, and pol*)[3] using derivational words sourced from UniMorph[4], which provides shallow morphological decomposition by segmenting only the outermost derivational morpheme. For example, $unresolvability$ is segmented as *un*<PREF> *resolvability*<N>, without further decomposing $resolvability$. To capture the full derivational structure, we extended the dataset by recursively decomposing words whenever additional derivational structure is present in the dataset, ensuring that all affixes are explicitly separated from the base form, as illustrated below:

```
unresolvability
un<PREF> resolvability<N>
un<PREF> re<PREF> solve<V> ability<N_SUFF>
un<PREF> re<PREF> solve<V> able<N> ity<N_SUFF>
un<PREF> resolve<V> able<N> ity<N_SUFF>
```

Each decomposition variant is preserved as an alternative for evaluation. Further preprocessing includes removing non-capitalized German nouns, multi-word expressions, and deduplication of lemmas to ensure no overlap between the few-shot training examples and test sets. Finally, words are categorized based on their part of speech: nouns(N), verbs(V), and adjectives(ADJ). See table 1.

Few-shot examples: We select N=[0, 3, 5, 10] examples with unique combinations of affix tags for each POS.

Prompt: We provide instructions that require the model to decompose derivational words without explanation and to assign part-of-speech tags selected from a predefined set. (See Appendix C for prompt instruction.)

3.3 Task 3: Morphological decomposition for Amharic and German

We introduce this task for Amharic and German, using a different data source from Task 2. For Amharic, only inflectional morphology is available in UniMorph. For German, not all analysis steps

[3]The five languages were selected from the set of ten based on the availability of derivational word lists in UniMorph.

[4]https://unimorph.github.io/

regarding derivation and compounding are covered, despite our efforts discussed in Section 3.2. This variant follows the same structure as Task 2, but enables a deeper investigation of derivational morphology in two typologically distinct languages, representing both high- and low-resource settings.

German Data The German dataset is created based on morphological analyses (obtained with SMOR (Schmid et al., 2004), which covers compounding, derivation, and inflection) of a large corpus (Wikipedia data). As the words are analyzed without sentence context and German has a high degree of syncretism, we only consider derivational analyses with unambiguous morphological decomposition and disregard the mostly ambiguous inflectional analyses. To obtain a clean dataset, we only use words with a non-ambiguous analysis at the level of word formation and apply some additional filtering steps[5]; see Appendix B.2 for more details.

From the resulting set, we randomly select nouns and adjectives[6] containing derivational operations (such as a `Suffix` tag). While the words are presented as inflected word forms, the analysis is in canonical form. We slightly modify SMOR's representation to obtain a sequence of morpheme-tag pairs, as illustrated below:

`Trocknungsvorgangs` (Noun)

`trocknen<V>`	`ung<NN_SUFF>`	`Vorgang<NN>`	`<+NN>`
dry_V	ing_{SUFF}	$process_N$	

While we focus on derivational analyses, compounding is very common in German word formation and is also reflected in our dataset. Obtaining a morphological analysis goes beyond splitting at the correct positions, but also requires non-concatenative processes, such as handling transitional elements (for example the *-s-* between *Trocknung* (*drying*) and *Vorgang* (*process*)) and mapping verb stems to lemmas (*trockn-* (*dry*) to *trocknen* (*to dry*)).

Amharic Data We use the WMT dataset[7] as our source corpus and analyze unique Amharic words using the HornMorpho morphological analyzer and generator tool[8]. Amharic exhibits a root-and-pattern morphology, a template-based system of word formation where roots, typically composed of three consonants, are combined with vocalic patterns to form stems. The root encodes the core lexical meaning, while the pattern provides grammatical and derivational information. We perform morphological analysis to obtain the full decomposition of each word into its constituent morphemes. For example, the verb **አልተሰበረም** - '*It was not broken*' - can be segmented as

አል-<ተ-ሰበር>-አ-ም

`NEG-PAS-break-3SG.M-NEG`,[9]

where <**ተ-ሰበር**> is a verb stem derived from the root **ሰብር**, and the prefix **ተ** is a passive/reflexive derivational morpheme. The full morpheme-level analysis is **አል**<ADV> **ተ**<V_PREF> **ሰበር**<VERB> **አ**<PRON> **ም**<ADV>.

To focus on derivational morphology, we eliminate inflectional variation of a lemma by selecting a single representative word per lemma, prioritizing outputs with a single valid decomposition. This lemma-level filtering ensures that there is no lemma overlap between training and test splits[10].

Few-shot examples and Prompt Task 3 follows the same approach as Task 2. However, in Task 2, we constrain POS tagging by classifying all morphemes appearing before the lemma as `PREFIX`, while suffixes are tagged based on the POS they derive, rather than their intrinsic linguistic form or function. (e.g., *-tion* in $information$ is labeled as `N_SUFF` and `inform` as V).

In contrast, Task 3 adopts a linguistically informed annotation scheme. Here, we annotate each morpheme with its actual linguistic POS, rather than categorizing it as a functional prefix or suffix. We restrict the derivational affixes for both languages to a set of well-attested, productive morphemes that also facilitate the reliable extraction of derivational words from the morphological analyzer output.

[5]SMOR provides several analyses, for example, with a varying level of granularity. As these variants can be equally plausible, we keep alternative analyses in some cases, such as *überschauen vs. über|schauen* (*overview*); cf. Appendix B.2

[6]We do not consider verbs in this experiment as they are less complex at the level of word formation.

[7]https://huggingface.co/datasets/allenai/wmt22_african

[8]https://github.com/hltdi/HornMorpho/tree/master

[9]Glosses: NEG = negative, PAS = passive, 3 = third person, SG = singular, M = masculine.

[10]Adjectives are excluded due to a narrow range of adjective forms in our data, which are often restricted to a single type and do not always function as adjectives. This reflects both limitations of the tool and the linguistic challenge in distinguishing Amharic adjectives from nouns, given their similar morphological formation.

4 Experiments and Results

For Task 1, we report the **micro F1 score** and for Tasks 2 and 3, we evaluate model performance by means of:

Tagging Accuracy: Exact match accuracy of the full predicted sequence, including both morphemes and their associated tags.

Segmentation Accuracy: Exact match accuracy based only on the morpheme sequence, ignoring tag correctness.

4.1 Labeling Syntactic Information

GPT-4o consistently achieves higher F1 scores across most languages and shot settings, with particularly strong performance in higher-resource languages such as *eng, deu,* and *ita*, as well as in low-resource languages like *amh* and *lav* (see Figure 1). In the case of *amh*, the test set contains relatively short sentences, which might have contributed to the comparatively high performance. While LLaMA shows greater gains for *amh* at the 3-shot setting, only minor improvement in GPT-4o. *amh* (in morph setting) remains challenging for both models.

GPT-4o exhibits inconsistency across N-shot settings in *eng, fra,* and *lit*. For LLaMA, performance in *amh, deu,* and *ita* shows minimal or no gains between 3- and 5-shot settings. Increasing few-shot examples does not help with *mlt*, though LLaMA performs best in the 3-shot setting.

LLaMA's overall weaker performance appears to stem primarily from difficulties in following instructions. Both models struggle with morphologically rich or low-resource languages such as *amh* (in the morph setting) and *lit* (note the limited test data cf. Table 2), though GPT-4o maintains a slight advantage in these cases.

Zero-shot performance is poor across all languages and models, with F1 scores below 30% (slightly higher F1 for *amh* and *eng* in GPT-4o), indicating models' limited prior knowledge for this task. Moving from $N = 0$ to $N = 1$ roughly doubles performance, but increasing further to $N = 3$ or $N = 5$ provides only marginal gains compared to $N = 1$.

4.2 Task 2: Morphological decomposition

Nouns As shown in Fig.2, *ita* nouns achieve the highest tagging accuracy in both models, with GPT-4o improving as N increases, followed by *eng*. Although higher N generally improves noun performance across languages, *lav* shows no gain from 3-shot to 5-shot. GPT-4o performs nearly equivalently in 0-shot and 3-shot for *deu, fra, pol*, and *lav*, whereas LLaMA shows clear gains.

Word segmentation accuracy is also highest for *fra* and *eng* (Fig.3). However, a significant decline (of 10%) in tagging accuracy underscores difficulties in assigning correct tags, particularly in GPT-4o. In contrast, *pol* nouns yield lower overall accuracy than other high-resource languages in both models, and poor segmentation performance as shown in Fig.3. This might also be correlated to the fact that *pol* exhibits comparatively high morphological complexity in our set of high-resourced languages. *deu* nouns show similar performance with GPT-4o and slightly higher accuracy in LLaMA. Nevertheless, both models struggle with segmenting and tagging *deu* nouns. Although tagging accuracy for *pol* and *lav* is more consistent across N compared to segmentation.

Adjectives Adjective performance varies across languages and models. Accuracy nearly doubles from zero-shot to 3-shot, but GPT-4o shows little to no improvement beyond $N > 3$, while LLaMA demonstrates consistent gains (Figure 2). Both models exhibit a slight decline in accuracy for English as N increases.

Tagging accuracy is generally lower for *deu, fra*, and *eng*, whereas *pol* and *lav* achieve comparatively better results. In *fra*, GPT-4o shows no improvement across N shots, while LLaMA yields significant gains. *deu* shows improvement at the 3-shot in GPT-4o, and *lav* benefits from higher N in GPT-4o, with LLaMA continuing to improve consistently across shots.

Verbs Similar to adjectives, performance improvement is shown from zero-shot to 3-shot. Increasing the number of shots further does not improve tagging performance for *pol* or *lav* verbs. *fra* and *deu* verbs are best at 5-shot and 3-shot, respectively, in both models. *eng* benefits from higher N in GPT-4o but shows only slight improvement at 5-shot. LLaMA struggles with *fra* verbs in segmentation and tagging.

4.3 Task 3: Morphological decomposition for *amh* and *deu*

For *amh*, both models show clear differences in handling nouns and verbs. Noun tagging benefits from higher N and models learn noun patterns;

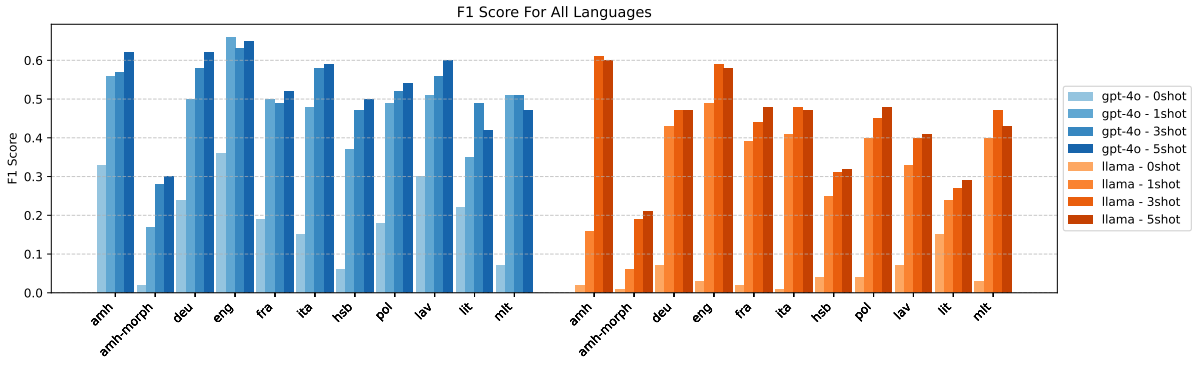

Figure 1: Labeling Syntactic Information (Task 1): **F1 score** across all languages.

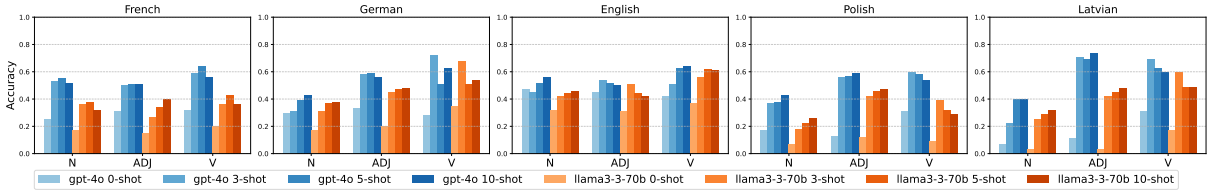

Figure 2: Morphological decomposition (Task 2): **Tagging Accuracy** for N-shot per language.

cf. Figures 4 and 5. Verbs, however, remain challenging across N shots. Even with increased shots, both models struggle to accurately tag and segment verbs, reflecting the greater morphological complexity of *amh* verbal forms. GPT-4o shows modest gains with additional examples, reaching an F1 score of 36% at 10-shot. LLaMA's improvements are minimal, with segmentation and tagging remaining weak at 10-shot.

For *deu*, both models show consistent improvements in tagging accuracy for adjectives and nouns as N increases, with the best performance at 10-shot. Both models perform well in tagging and segmentation for adjectives. However, *deu* nouns, for both segmentation and tagging performance, lag behind adjectives. Tagging accuracy is particularly impacted by segmentation errors, as incorrect morpheme boundaries lead to incorrect tag assignments. The most substantial improvement is observed in LLaMA, with a 23% gain in tagging accuracy from 5-shot to 10-shot. Similarly, GPT-4o shows its highest gain at 10-shot as well, with a 13% improvement.

Overall, the results suggest that both models may lack basic morphological knowledge of *amh*, as nouns and verbs score zero or near zero in the zero-shot setting (more in Section 5). In contrast, for *deu*, the models mainly struggle with correctly tagging noun and adjective morphemes.

In addition, we investigated the effect of the prompt language, contrasting English, which might benefit from the model's stronger English representation (Zhao et al., 2024), with a human-translated and a machine-translated prompt (obtained through *Google translate*) in the respective target languages *deu* and *amh* (Fig. 8). We found little performance differences across these variants, suggesting that even for a low-resource language like Amharic, a machine-translated prompt can be as successful as a human-translated one. However, this might not be representative since we examine only one task, and the label set used in this experiment is in English.

5 Error analysis and Discussion

5.1 Labeling Syntactic Information

Error patterns across GPT-4o and LLaMA reveal shared trends in argument identification, with notable differences in how performance scales with increasing shots.

For **subject** identification, both models perform well, aided by the high frequency of subjects and their typical placement at the beginning of sentences in SOV languages.

Object identification shows only moderate improvement in both models. GPT-4o benefits more clearly from additional examples, gradually reducing mislabeling. In low-resource languages like *amh* (morph setting) and *lit*, both models show limited gains, suggesting that data sparsity constrains

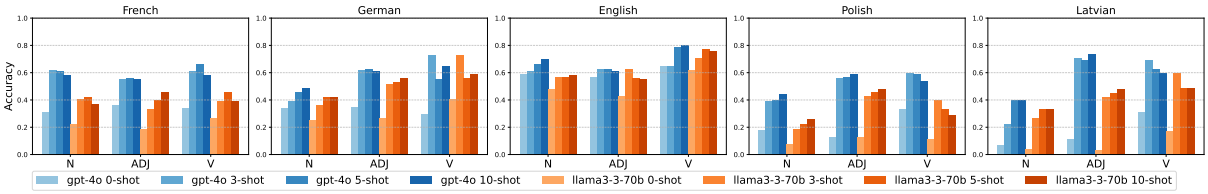

Figure 3: Morphological decomposition (Task 2): **Segmentation Accuracy** for N-shot per language.

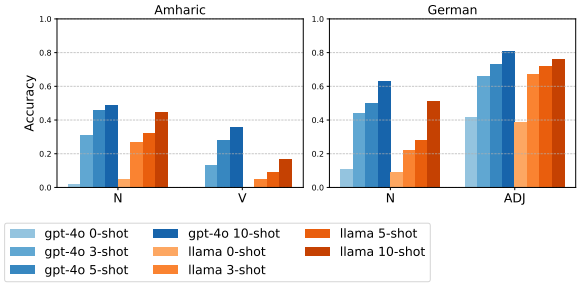

Figure 4: Morphological decomposition (Task 3): **Tagging Accuracy** for N-shot for **Amharic** and **German**.

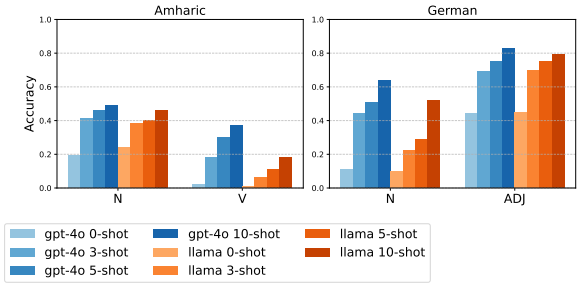

Figure 5: Morphological decomposition (Task 3): **Segmentation Accuracy** for N-shot for **Amharic** and **German**.

object detection.

Indirect object identification remains the most challenging task for both models. While GPT-4o demonstrates slightly better consistency, neither model shows substantial improvement as the N increases. In *amh* and *pol*, LLaMA's performance deteriorates with higher N, exhibiting more errors at 3- and 5-shot. This difficulty appears across languages, likely due to the under-representation of indirect objects in the dataset, resulting in the models seeing too few examples in the few-shot settings.

5.2 Morphological Decomposition

Task 2 The model performs well overall in labeling POS tags across languages. However, when prompted with a specific tag, it tends to over-predict that category or high-frequency morphological tags (e.g, prefix). For example, when "noun" is specified, the model often over-predicts nouns and noun suffixes; similarly, when the main tag is specified as "verb", there is a tendency to over-predict verbs and verb suffixes. Additionally, LLaMA exhibits issues with instruction following: it frequently provides explanations even when explicitly prompted not to, and fails to consistently adhere to the output template specified in the few-shot examples.

Task 3 Both models showed greater difficulty with segmentation than with tagging. We take a closer look at segmentation errors on the best performance (10 shots). For *amh*, we examined the top 10 most frequently missed morphemes, which accounted for at least 20% of the total segmentation errors. Most of these morphemes are bound morphemes, which present a challenge in a language like *amh* that follows a root-and-pattern morphology. Moreover, accurate segmentation often requires separating vowels from consonants, which are represented by a single character, a task that both models frequently struggle with.

As illustrated in the example below, the models fail to segment words with non-linear segmentation.

word: ካላሳቀህ (*"If it hasn't made you[masc] laugh"*)

correct: ከ(if) አል(not) አስ(aspect.PFV) ሳቅ(laugh) አ(it) ህ(2;M;SG)

LLaMA: ካ - አል - ሰቀህ

GPT-4o: ካ - ላ- አ - ሳቅ - እህ

The main difficulty lies in how these morphemes fuse in surface realization. For instance, when አስ is placed between አል and ሳቅ, the ል in አል shifts to ላ. Similarly, the verb stem ሳቅ undergoes morphophonemic alternation when followed by the third-person marker አ, whereby the final consonant ቅ shifts to ቀ. GPT-4o successfully identified the verb stem but reduced the prefixes to their surface realizations instead of their underlying morphemes. LLaMA, by contrast, captured the negation morpheme but failed to recover the verb stem.

We analyzed how LLaMA and GPT-4o handle noun and verb bound morphemes for *amh*, focusing

on segmentation errors. Both models struggle with the same types of bound morphemes, but GPT-4o makes roughly half as many errors.

We analyze *deu* nouns and adjectives to identify where the models struggle and which morphological patterns are difficult. A key challenge is correctly transforming verbs into their canonical forms, which typically end with *-en* or *-n*, as illustrated in the following example (analysis with GPT-4o):

word: *Größenänderungen* (*"size*$_N$ *change*$_V$ *ing*$_{N\_SUFF}$*": changes in size*)

correct: *Größe*<NN> *ändern*<V> *ung*<NN_SUFF>

GPT-4o: *groß*<ADJ> *en*<ADJ_SUFF> *änder*<V> *ung*<NN_SUFF>

The verb *ändern* (*to change*) requires an *-n* at the end to mark the infinitive form. In GPT-4o's output, this *-n* is missing, such that the analysis only contains the verb stem *änder-*. In addition, the model did not recognize the noun *Größe* (*size*), but instead output an incorrect decomposition into an adjective *groß* and a suffix[11].

We examined model outputs, disregarding the correctness of word segmentation, to assess the models' underlying understanding of *deu* word classes. We observe that lower N-shot pose challenges for recovering the correct canonical form, particularly in the LLaMA model. Increasing N doubled the rate of correct verb transformations in both models.

Looking into structural formulation, we extracted tag sequences only and computed BLEU scores over these sequences (see Fig. 7). The findings align with those shown in Fig. 4, suggesting that models are capable of learning abstract grammatical patterns. Notably, the gains in LLaMA models between 3- and 5-shot were marginal, compared to Fig. 4, possibly due to limitations in handling *deu* verb morphology.

Furthermore, our evaluation did not consider case-based orthographic rules. In standard *deu*, nouns are capitalized while verbs are typically lowercase. To better understand whether the model encodes this rule, we analyzed errors related to the capitalization of nouns and verbs. We found that LLaMA showed a higher noun capitalization error rate in the 3- and 5-shot, while both models exhibited minimal errors in verb capitalization.

[11] *-en* can be an inflectional adjective suffix, but not in this context. There is a transitional element *-n-* between the nouns *Größe* and *Änderung* that should, however, not be part of the analysis as it carries no meaning.

Finally, we examine partial segmentation accuracy (cf. Fig. 6), which evaluates how well constituent morphemes are segmented across all N-shots. Scores are consistently higher for both languages and across all N-shots, indicating that models can correctly identify some morphemes. Nevertheless, partial accuracy alone is insufficient, as proper segmentation requires all morphemes. Thus, while the models capture some surface regularities, this does not necessarily imply deeper morphological understanding.

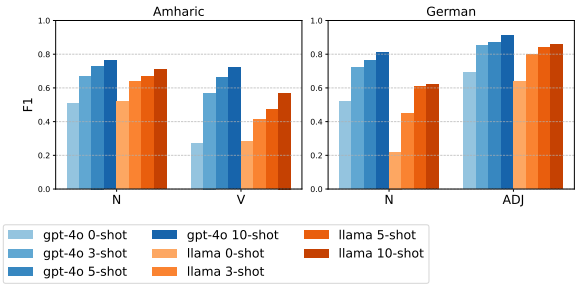

Figure 6: Morphological decomposition (Task 3): **Partial Segmentation Accuracy**

6 Conclusion

This study explored methods to extract linguistic information from large language models (GPT-4o and LLaMA 3.3-70B) covering a wide range of languages. We conducted three tasks: (i) labeling syntactic information (at the sentence level), (ii) morphological decomposition of derived words (at the word level), and (iii) a detailed analysis of morphological derivation for German and Amharic.

Our findings show that GPT-4o consistently demonstrated stronger linguistic knowledge than LLaMA across all tasks. However, both models exhibit clear limitations in acquiring robust linguistic knowledge, with challenges observed across all languages and tasks. While labeling syntactic information benefits from higher N-shot, morphological decomposition remains particularly difficult; scores remain low even for high-resource languages where stronger performance would be expected. To better understand these limitations, we conducted a detailed analysis for *amh* and *deu*.

Overall, our results align with the findings of Di Marco and Fraser (2024), Kodner et al. (2023), and Ismayilzada et al. (2025) that while the models are capable of capturing abstract patterns, this does not necessarily translate into a robust understanding or generalization of the morphosyntactic behavior of languages.

7 Limitations

Languages and Models While we consider several typologically diverse languages, covering both high- and low-resource scenarios, they are by no means representative of the entire diversity range of languages. Furthermore, we only investigated a subset of morphological operations, namely derivational morphology for nouns, verbs, and adjectives. This is largely due to the availability of datasets providing fine-grained and consistent derivational analyses across different languages.

Similarly, considering more models might provide further insights into the abilities of different model families, in particular with regard to the coverage of different languages.

Dataset The dataset used in Task 1 contains some annotation errors and shows a clear imbalance across the syntactic categorise we aim to identify. Likewise, Task 2 does not account for all possible derivations of a word. We attempted to derive the smallest decomposition of words by referencing examples within the corpus, which does not always result in canonical word forms. In addition, not all words in this dataset are derivational. In effort to improve this analysis, we used finite-state transducer (FST) tools for German and Amharic. However, the FST analyzers have their own limitations, including instances of over- or under-segmentation, and variability in analysis where words may have multiple interpretations.

Subword segmentation The task of morphological decomposition is directly related to the underlying subword segmentation: while subwords can provide access to word parts, most subword segmentation strategies are lingusitically uninformed and the resulting subwords thus do not necessarily correspond to linguistically meaningful units. There is a large body of research on subword segmentation, in particular with regard to the representation of low-resource languages and languages of other scripts in English-dominated LLMs, that generally reaches the conclusion that linguistically inspired approaches are beneficial (for example, Hofmann et al. (2021); Hou et al. (2023); Limisiewicz et al. (2024)).

While the representation of subwords are undeniably a relevant aspect for obtaining derivational analyses, the exploration of the impact of subword segmentation is beyond the scope of this work.

Prompting Language and Terminology Our experiment on contrasting prompt languages showed that similar results could be obtained with English, human- and machine-translated prompts. This finding is particularly important for low-resource languages, where machine translation can serve as an alternative in the absence of human translators. In our study, however, only the prompts were translated, while the tags remained in English, leaving unexplored the potential impact of using language-specific terminology on model performance.

8 Acknowledgements

The work was supported by the European Research Council (ERC) under the European Unions Horizon Europe research and innovation programme (grant agreement No. 101113091) and by the German Research Foundation (DFG; grant FR 2829/7-1).

References

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.

Jiangang Bai, Yujing Wang, Yiren Chen, Yaming Yang, Jing Bai, Jing Yu, and Yunhai Tong. 2021. Syntax-bert: Improving pre-trained transformers with syntax trees. *Preprint*, arXiv:2103.04350.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2023. Prompting language models for linguistic structure. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6649–6663, Toronto, Canada. 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, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Marion Di Marco and Alexander Fraser. 2024. Subword segmentation in LLMs: Looking at inflection and consistency. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12050–12060, Miami, Florida, USA. Association for Computational Linguistics.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Linyang He, Peili Chen, Ercong Nie, Yuanning Li, and Jonathan R. Brennan. 2024. Decoding probing: Revealing internal linguistic structures in neural language models using minimal pairs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4488–4497, Torino, Italia. ELRA and ICCL.

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.

Jue Hou, Anisia Katinskaia, Anh-Duc Vu, and Roman Yangarber. 2023. Effects of sub-word segmentation on performance of transformer language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7413–7425, Singapore. Association for Computational Linguistics.

Mete Ismayilzada, Defne Circi, Jonne Sälevä, Hale Sirin, Abdullatif Köksal, Bhuwan Dhingra, Antoine Bosselut, Duygu Ataman, and Lonneke van der Plas. 2025. Evaluating morphological compositional generalization in large language models. *Preprint*, arXiv:2410.12656.

Jordan Kodner, Sarah Payne, Salam Khalifa, and Zoey Liu. 2023. Morphological inflection: A reality check. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6082–6101, Toronto, Canada. Association for Computational Linguistics.

Zhongli Li, Qingyu Zhou, Chao Li, Ke Xu, and Yunbo Cao. 2021. Improving BERT with syntax-aware local attention. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 645–653, Online. Association for Computational Linguistics.

Tomasz Limisiewicz, Terra Blevins, Hila Gonen, Orevaoghene Ahia, and Luke Zettlemoyer. 2024. MYTE: Morphology-driven byte encoding for better and fairer multilingual language modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15059–15076, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. *Preprint*, arXiv:2101.05779.

Rita Ramos, Everlyn Asiko Chimoto, Maartje ter Hoeve, and Natalie Schluter. 2024. Grammamt: Improving machine translation with grammar-informed in-context learning. *Preprint*, arXiv:2410.18702.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. SMOR: A German computational morphology covering derivation, composition and inflection. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book. *Preprint*, arXiv:2309.16575.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1331 others. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Andreas Waldis, Yotam Perlitz, Leshem Choshen, Yufang Hou, and Iryna Gurevych. 2024. Holmes: A benchmark to assess the linguistic competence of language models. *Transactions of the Association for Computational Linguistics*, 12:1616–1647.

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.

Marion Weller-Di Marco and Alexander Fraser. 2024. Analyzing the understanding of morphologically complex words in large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1009–1020, Torino, Italia. ELRA and ICCL.

Zenan Xu, Daya Guo, Duyu Tang, Qinliang Su, Linjun Shou, Ming Gong, Wanjun Zhong, Xiaojun Quan, Daxin Jiang, and Nan Duan. 2021. Syntax-enhanced pre-trained model. 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 5412–5422, Online. Association for Computational Linguistics.

Chen Yang. 2021. Learning better sentence representation with syntax information. *CoRR*, abs/2101.03343.

Kexun Zhang, Yee Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024. Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15654–15669, Bangkok, Thailand. Association for Computational Linguistics.

Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llama beyond english: An empirical study on language capability transfer. *Preprint*, arXiv:2401.01055.

A Appendix

A.1 UD Dataset

We conduct our experiments on ten languages sourced from the UD dataset. Amharic (UD_Amharic-ATT[12]), English (UD_English-EWT[13]), French (UD_French-Sequoia[14]), German (UD_German-GSD[15]), Italian (UD_Italian-ISDT[16]), Latvian (UD_Latvian-LVTB[17]), Lithuanian (UD_Lithuanian-HSE[18]), Maltese (UD_Maltese-MUDT[19]), Polish (UD_Polish-PDB[20]), Upper Sorbian (UD_Upper_Sorbian-UFAL[21]).

A.2 Language Properties

Amharic is a member of the Afro-Asiatic language family, belonging to the Semitic branch. It uses the Ge'ez script. Amharic is a pro-drop language and exhibits templatic and concatenative morphemes. Templatic morphology, highly productive across Semitic languages, is organized around root morphemes - typically that convey core semantic concepts. Syntactically, Amharic default word order is Subject-Object-Verb (SOV). Verbs are primarily derived from verbal roots and inflected using a combination of prefixes and suffixes to mark subject agreement. While nouns and adjectives are commonly derived from verbal roots, adjectives can also be derived from nouns. Regarding adverbs, Amharic possesses a small set of monomorphemic, underived adverbs and conjunctions, though most adverbial and conjunctive expressions are formed from nouns with prepositional prefixes and, sometimes, postpositions.

English belongs to the Germanic branch of the Indo-European language family. It is a moderately analytic language that follows a subject-verb-object (SVO) word order.

German is a member of the Germanic branch of the Indo-European language family and is spoken by over 130 million people. It is a fusional language and its word order follows an SVO (subject-verb-object) structure, while allowing for some flexibility. German has a rich nominal morphology, inflecting for case, number, gender and strong/weak inflection. However, it also exhibits a high degree of syncretism, such that an observed word form often cannot be analyzed for these features without sentence context. German has very productive word formation processes, including derivation and compounding.

French is a Romance language - a part of the Indo-European language family. It uses Latin script with four diacritics appearing on vowels and fol-

[12] https://github.com/UniversalDependencies/UD_Amharic-ATT/tree/master

[13] https://github.com/UniversalDependencies/UD_English-EWT/tree/master

[14] https://github.com/UniversalDependencies/UD_French-Sequoia/tree/master

[15] https://github.com/UniversalDependencies/UD_German-GSD/tree/master

[16] https://github.com/UniversalDependencies/UD_Italian-ISDT/tree/master

[17] https://github.com/UniversalDependencies/UD_Latvian-LVTB/tree/master

[18] https://github.com/UniversalDependencies/UD_Lithuanian-HSE/tree/master

[19] https://github.com/UniversalDependencies/UD_Maltese-MUDT/tree/master

[20] https://github.com/UniversalDependencies/UD_Polish-PDB/tree/master

[21] https://github.com/UniversalDependencies/UD_Upper_Sorbian-UFAL/tree/master

lows subject-verb-object (SVO) word order.

Italian is a Romance language and part of the Indo-European language family. It uses the Latin script and shares a high lexical similarity with French. Italian has flexible word order and often omits the subject, which is typically indicated by verbal inflections.

Polish is a Slavic language that belongs to the Indo-European language family. Its alphabet is based on the Latin script but includes extra letters with diacritical marks. Polish is a highly fusional language and follows a typical subject-verb-object (SVO) structure, although word order can be relatively flexible. There are no articles, and subject pronouns are often dropped.

Latvian is an East Baltic language within the Indo-European language family. It uses the Latin script; the basic word order in Latvian is subject-verb-object; however, the word order is relatively free.

Lithuanian is an East Baltic language within the Indo-European language family. The language uses the Latin alphabet and is characterized by a high degree of inflection. Lithuanian also features an extensive system of word formation, which contributes to its lexical richness.

Maltese is a Semitic language belonging to the Afro-Asiatic language family. It is primarily written in the Latin script. While the typical word order is subject-verb-object (SVO), Maltese allows for considerable flexibility in sentence structure.

Upper Sorbian is a Slavic language within the Indo-European language family. It is written using the Latin script and typically follows a subject-verb-object (SVO) word order.

B Datasets

B.1 Amharic dataset Details (Task 1)

For example, subject agreement morphemes, which encode information about the subject of the verb, are often segmented from the verb and annotated as separate syntactic words in the UD dataset. However, these elements function more like bound morphemes: they cannot occur independently and often appear alongside an overt subject.

- Sentence:
  በመናገር ላይ እያለሁ መልስ ሰጠችኝ።
  While speaking I was an answer she gave me.
  (She gave me an answer while I was speaking.)
- Word label:
  በመናገር ላይ እያለሁ [መልስ | object] [ሰጠችኝ | indirect_object | subject] ።
- Morpheme label:
  በመናገር ላይ እያለሁ [መልስ | object] ሰጥ[አች(she) | subject][ኝ(me) | indirect_object] ።

In the example, the verb **"ሰጠችኝ"** (*she gave me*) simultaneously encodes both the subject and the indirect object within a single word. This creates two possible labeling strategies: (i) labeling the entire word with both syntactic roles, or (ii) segmenting the word and labeling the individual morphemes that indicate subject and indirect object.

To capture this distinction, we prepared two versions of the Amharic UD dataset. The first, referred to as **Amharic**(*amh*), labels words that bear core syntactic roles. The second, called **Amharic-morph** (*amh-morph*), uses a morpheme-level annotation approach, as illustrated in the morpheme label example.

B.2 German Dataset Details (Task 3)

Due to its finite-state architecture, SMOR provides all possible analyses of an input word. While it does not output ranking criteria such as probabilities, there is a more restrictive setting that excludes less plausible analyses. We base our dataset on analyses using this setting, while additionally restricting that the analyses remain ambiguous after removing inflectional features.

One persisting problem is the level of granularity: while SMOR generally can provide very fine-grained analyses, they are sometimes prevented in the restricted setting. We thus consider alternative evaluations to allow for different levels of granularity, for example, with regard to the splitting of (lexicalized) compounds or the splitting of prefixes/particles in verbs, such as *überschauen* vs. *über|schauen* (*to over|view*).

Setting the segmentation granularity to a "universally good level" is difficult, as this can depend on the actual words, but might also vary with different types of downstream applications. Being primarily interested in evaluating the LLMs' ability to generate *a* plausible analysis, allowing for alternative solutions for a particular set of words, is a straightforward way to accommodate this issue.

B.3 Details on Few-shot Selection (Task 1)

For our experiments in Task 1 – identifying subjects, objects, and indirect objects across ten diverse languages, we curated a test set aimed at representing four key syntactic categories: subject only, subject and object, subject and indirect object, and all. Our initial plan was to select a total of

500 sentences, ideally with 125 from each category. However, not all languages in our collection contained data for every category. Therefore, instead of enforcing a uniform distribution, we preserved the naturally unbalanced dataset, as we believe it more accurately reflects real-world language data as found in available corpora (Refer to Table 2).

Code	Language	POS Tag	Test set count
fra	French	ADJ	1000
		N	1000
		V	1000
deu	German	ADJ	1000
		N	1000
		V	1000
eng	English	ADJ	1000
		N	1000
		V	1000
pol	Polish	ADJ	1000
		N	1000
		V	1000
lav	Latvian	ADJ	62
		N	300
		V	35

Table 1: Morphological decomposition (Task 2) - number of test sets per language.

Language	**1_shot**	**3_shot**	**5_shot**	**Test**
amh	3	9	14	251
eng	4	12	20	285
deu	4	12	20	311
fra	4	12	20	263
ita	4	12	20	258
hsb	3	7	10	256
pol	4	12	20	411
lav	4	12	20	465
lit	4	12	18	56
mlt	4	12	18	225

Table 2: Number of sentences selected in N-shot and test set (Task 1)

C Instructions

We develop English prompt instructions for three morphosyntactic tasks, adapting the template from Paolini et al. (2021). The first task centers on labeling syntactic information (see Table 3), while the second addresses morphological decomposition in five typologically diverse languages (see Table 4). Furthermore, we provide a detailed study of morphological decomposition in Amharic and German (see Tables 5, 6, and 7).

For the zero-shot setting, we add extra instructions to guide the model toward the expected output format. For the N-shot setting (N > 0), however, we rely on the model to learn the format directly from the provided examples.

D BLUE score for German Morphological Tag Sequence

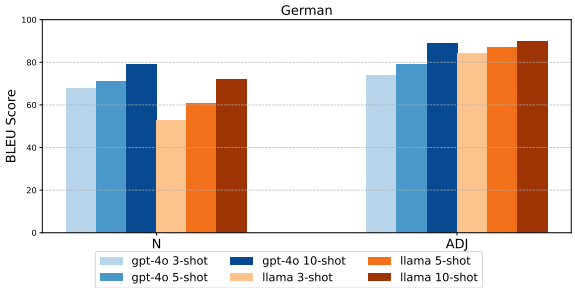

Figure 7: Blue Score for German Morphological Tag Sequence

E Result for In-language Prompt Instructions

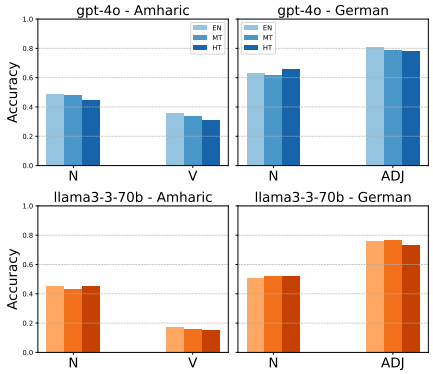

Figure 8: Morphological decomposition (Task 3): **Tagging Accuracy** on instructions in English(EN), Machine Translated(MT) and Human Translated(HT).

Language	Instruction	Example
English (zero-shot)	You are a linguistic analyser. You will receive a sentence with exactly one verb enclosed in [ ]. Label only that verbs subject as [subject], its object as [object], and its indirect object as [indirect_object] if present. The verb should not be labeled in your output. Maintain the original sentence wording, grammar, and structure. Only append the appropriate label directly after the correct word in the sentence. If a subject, object, or indirect object is a multi-word phrase, label only the rightmost word of the phrase. Output the labeled sentence without additional explanation. The input and output sentence should follow the following format. Input sentence: Question: WORD-1 WORD-2 [WORD-3] WORD-4 WORD-5. Output sentence: WORD-1 [WORD-2 | subject] WORD-3 [WORD-4 | object] WORD-5.	"question": "Please feel free to give me a call if you [have] any questions concerning the attached guaranties ." "expected_answer": "Please feel free to give me a call if [you | subject] have any [questions | object] concerning the attached guaranties ." "model_output": "Please feel free to give [me | indirect_object] a call if [you | subject] have any questions concerning the attached guaranties."
English	You are a linguistic analyser. You will receive a sentence with exactly one verb enclosed in [ ]. Label only that verbs subject as [subject], its object as [object], and its indirect object as [indirect_object] if present. Maintain the original sentence wording, grammar, and structure. Only append the appropriate label directly after the correct word in the sentence. If a subject, object, or indirect object is a multi-word phrase, label only the rightmost word of the phrase. Output the labeled sentence without additional explanation.	"question": "They [walked] me through all the steps involved in the installation project so that there were no surprises ." "expected_answer": "[They | subject] walked [me | object] through all the steps involved in the installation project so that there were no surprises ." "model_output": "[They | subject] walked [me | object] through all the steps involved in the installation project so that there were no surprises ."

Table 3: Instruction for Labeling Syntactic Information (Task 1)

Language	Instruction
English	You are a morphological analyser. Find the derivational analysis of the given [$lang$] [tag] word. Use only the predefined morphological tags. Lemmatize each word and ignore inflectional morphemes. You only need to provide the morphological analysis with no further explanations. Tags: <N> for a noun. <ADJ> for an adjective. <ADV> for an adverb. <V> for a verb. <ADJ_SUFF> for adjective suffix. <N_SUFF> for noun suffix. <ADV_SUFF> for adverb suffix. <V_SUFF> for verb suffix. <PREF> for all prefixes.

Table 4: Instruction for Morphological decomposition task (Task 2)

Language	Instruction
German	You are a morphological analyser. You are given a morphologically complex [$lang$] [tag]. Your task is to output the derivational analysis of that word. Use only the predefined morphological tags listed below. All components should be lemmatized. You only need to provide the morphological analysis with no further explanations. Tags: <NN> for a noun. <ADJ> for an adjective. <VPART> for a verb particle. <V> for a verb. <PREF> for prefix. <ADJ_SUFF> for adjective suffix. <NN_SUFF> for noun suffix. **(for zero shot experiment we add the following instruction)** Output the labeled sentence without additional explanation. The input and output sentence should follow the following format. Input sentence: Question: WORD Output sentence: morpheme-1<Tag> morpheme-2<Tag> ...
Amharic	You are a morphological analyser. You are given a morphologically complex [$lang$] [tag]. Your task is to output the derivational analysis of the word and provide its smallest possible morphological decomposition. Use only the predefined morphological tags listed below. You only need to provide the morphological analysis with no further explanations. Tags: <N> for a noun. <N_PREF> for a noun prefix. <N_SUFF> for noun suffix. <ADJ> for an adjective. <ADV> for an adverb. <PRON> for a proper noun. <PART> for a verb particle. <V> for a verb. <V_PREF> for verb prefix. <AUX> for an auxiliary verb. <ADP> for Adposition. **(for zero shot experiment we add the following instruction)** Output the labeled sentence without additional explanation. The input and output sentence should follow the following format. Input sentence: Question: WORD Output sentence: morpheme-1<Tag> morpheme-2<Tag> ...

Table 5: Instruction for Morphological decomposition task (Task 3)

Language	Instruction
	Machine Translate
German	Sie sind ein morphologischer Analysator. Ihnen wird ein morphologisch komplexes deutsches [tag] vorgelegt. Ihre Aufgabe ist es, die Ableitungsanalyse dieses Wortes auszugeben. Verwenden Sie ausschlieSSlich die unten aufgeführten vordefinierten morphologischen Tags. Alle Komponenten sollten lemmatisiert sein. Sie müssen lediglich die morphologische Analyse ohne weitere Erläuterungen angeben. Tags: <NN> für ein Nomen. <ADJ> für ein Adjektiv. <VPART> für einen Verbpartikel. <V> für ein Verb. <PREF> für ein Präfix. <ADJ_SUFF> für ein Adjektivsuffix. <NN_SUFF> für ein Nomensuffix.
Amharic	እርስዎ የሞሮሎጂ ተንታኝ ነዎት። ሞርሮሎጂያዊ ውስብስብ አማርኛ [tag] ተሰጥቶሃል። የእርስዎ ተግባር የቃሉን የመነሻ ትንተና ማውጣት እና አነስተኛውን የሞርፎሎጂ መበስበስ ማቅረብ ነው። ከዚህ በታች የተዘረዘሩትን አስቀድሞ የተገለጹ የሞርፎሎጂ መለያዎችን ብቻ ይጠቀሙ። ምንም ተጨማሪ ማብራሪያ ሳይኖር የሞርፎሎጂያዊ ትንታኔን ብቻ መስጠት ያስፈልግዎታል። መለያዎች <N> ለስም። <N_PREF> ለስም ቅድመ ቅጥያ። <N_SUFF> ለስም ቅጥያ። <ADJ> ለቅጽል። <ADV> ለተውላጠ ቃል። <PRON> ለትክክለኛ ስም። <PART> ለግስ ቅንጣት። <V> ለግስ። <V_PREF> ለግስ ቅድመ ቅጥያ። <AUX> ለረዳት ግስ። <ADP> ለማስታወቂያ።

Table 6: Instruction for Morphological decomposition task (Task 3): Translated prompts in Table 5 using *Google translate*

Language	Instruction
	Human Translate
German	Du erstellst linguistische Analysen. Finde die Derivation des gegebenen deutschen [tag]. Benutze nur die vorgegebenen morphologischen Tags. Lemmatisiere jedes Wort und ignoriere Flexionsmorpheme. Gib nur die morphologische Analyse aus, ohne weitere Erklärungen. Tags: <NN> für Nomen. <ADJ> für Adjektive. <VPART> für Verbpartikeln. <V> für Verben. <PREF> für Präfixe. <ADJ_SUFF> für Adjektivsuffixe. <NN_SUFF> für Nominalsuffixe.
Amharic	እርሶ የሥነቃላት አወቃቀር ባልሙያ ነዎት። ውስብስብ የአማርኛ [tag] ተሰጦታል። ከእርሶ የሚጠበቀው የቃሉን አመሰራረት በመከፋፈል የቃሉን መሰራች ምዕላዶች ያውጡ። ከታች የተጠቀሰውን የቃላት ክፍሎች ብቻ ይጠቀሙ። ያለምንም ተጨማሪ ማብራሪያ የቃሉን ምዕላዶች ብቻ ያውጡ። የቃላት ክፍሎች <N> ለስም። <N_PREF> ለስም ቅድመ ቅጥያ። <N_SUFF> ለስም ቅጥያ። <ADJ> ለቅጽል። <ADV> ለተውሳክ ግስ። <PRON> ተውላጠ ስም። <PART> ለመስተአምር። <V> ለግስ። <V_PREF> ለግስ ቅድመ ቅጥያ። <AUX> ለረዳት ግስ። <ADP> ለመስተዋድድ።

Table 7: Instruction for Morphological decomposition task (Task 3): Human Translated the prompt in Table 5