Analyzing the Understanding of Morphologically Complex Words in Large Language Models

Marion Weller-Di Marco[1,2] **and Alexander Fraser**[1,2,3]
[1]School of Computation, Information and Technology, TUM
[2]Center for Information and Language Processing, LMU Munich
[3]Munich Center for Machine Learning
marion.dimarco@tum.de

Abstract

We empirically study the ability of a Large Language Model (*gpt-3.5-turbo-instruct*) to understand morphologically complex words. In our experiments, we looked at a variety of tasks to analyse German compounds with regard to compositional word formation and derivation, such as identifying the head noun of existing and novel compounds, identifying the shared verb stem between two words, or recognizing words constructed with inappropriately used derivation morphemes as invalid. Our results show that the language model is generally capable of solving most tasks, except for the task of identifying ill-formed word forms. While the model demonstrated a good overall understanding of complex words and their word-internal structure, the results also suggest that there is no formal knowledge of derivational rules, but rather an interpretation of the observed word parts to derive the meaning of a word.

Keywords: Morphology, Linguistic information in LLMs, Compounds

1. Introduction

Large Language Models (LLMs) have been getting continually better at a large variety of tasks, and there is a growing interest in their linguistic abilities. It has been shown that pre-trained language models (PLMs) encode different types of linguistic information, for example syntactic knowledge (e.g. Tenney et al. (2019), Liu et al. (2019)), morphosyntactic properties (e.g. probing for grammatical number (Lasri et al., 2022)), multilingual probing for inflectional features (Shapiro et al., 2021)), semantic information (e.g. Ettinger (2020)) and world knowledge (e.g. Petroni et al. (2019)).

Morphologically complex languages are challenging for NLP, as a large amount of information is condensed into a single word, unlike in analytical languages where separate words make it easier to derive meaning. Morphologically complex words can be constructed through processes such as derivation and compounding, where word forms are created by adding prefixes and/or suffixes following language-specific patterns (derivation) and the concatenation of words (compounding). For example, the German verb *nutzen* (*to use*) can be derived into the adjective *nutzbar* (*usable*), and the noun *Nutzbarkeit* (*usability*). We can continue to create more complex words by means of compounding: *Nutzbarkeitsdauer, Gesamtnutzbarkeitsdauer, Gesamtnutzbarkeitsdauerstudie, ...* (*usability time, total usability time, total usability time study, ...*). These are productive word formation processes that often result in novel or infrequent words, in particular when also taking into account inflectional variants, thus making the ability to interprete complex and potentially novel words crucial to model languages that apply these processes.

There are numerous studies investigating linguistic properties and capabilities of LLMs, but in general with a strong focus on English; and, possibly as a consequence of English being a morphologically poor language, there is not much work addressing word formation or derivation in LLMs. As there are many languages with complex words, the ability to understand complex words and their word-internal structure is an important part in accommodating the processing of languages with a richer morphology.

While there is agreement that LLMs encode information about the relation between words (both at the syntactic and semantic level), it is not clear whether this also holds for the components of complex words. Considering complex words as syntactic structures in the sense that they are created from smaller units according to a defined set of rules (e.g. derivation patterns), we empirically study to what extent LLMs are capable of understanding the structure of complex words, by means of asking the model for particular parts of the analysis. For example, consider the following German nouns that share the substring *-plan-*:

- *Aushilfs|ka**plan*** (*temporary chaplain*)
- *Kinder|**plan**sch|becken* (*kids' paddle pool*)
- *Bebauungs|**plan*** (*development plan*)
- *Städte|**plan**er* (*urban planner*)
- ***Plan**ungs|dauer* (*planning duration*)

An obvious task is the identification of the head: *plan* for the word *Bebauungsplan*, but *kaplan* (*chaplain*) for *Aushilfskaplan*. With view to generalization, a LM likely benefits from knowing whether

LREC-COLING 2024, pages 1009–1020
20-25 May, 2024.

words are related: while the words all contain the substring *-plan-*, only the last three share the common meaning of *planen* (*to plan*), whereas the first two words are entirely unrelated. Structural understanding and knowledge about compounding, derivation patterns (e.g. nominalization) and morpheme boundaries help to uncover such relations.

In our experiments, we focus on grammatical analysis rather than semantic interpretation. The tasks are designed such that they ask for the identification of a particular part of a complex word or its decomposition, as well as narrow-framed generation tasks of word variations. We focus on German which has productive word formation and derivation processes. Furthermore, as a mid-high resource language, it is reasonably well represented in the LLM pre-training data. The experiments also aim at assessing whether explicit linguistic information can be accessed, or whether the model rather relies on it indirectly, by looking at whether the output is a direct answer to the question, or just demonstrates a general understanding of the target word, but not the question.

We designed a variety of tasks comprising compositional word formation and derivation processes, for example identifying the head noun of existing and novel compounds, the shared verb stem between two words, generating a variant of a given word (e.g. the non-negated form), or recognizing words constructed with inappropriately used derivation morphemes as invalid.

Our results show that the LM is generally capable of solving tasks that ask about particular parts of a word, although not perfectly. Furthermore, our results indicate that the model tends to understand the relation between the components of a word, even if the answer is not always correct in the context of the question, for example by defining a compound's meaning instead of giving the head noun. In contrast, the model largely failed to identify incorrect word forms, suggesting that there is no formal knowledge of derivational rules, but rather that the model relies on the interpretation of the individual parts to derive the meaning of a word.

2. Related Work

Morphology has been a central part in linguistic research for a long time, and has lead to numerous strands of research. One line of work conceptually related to some of our experiments is morphological re-inflection (e.g., the series of the respective Shared Tasks starting with Cotterell et al. (2016a)), where the inflected form for a given pair of lemma and morphological tag has to be generated. Another area is that of word segmentation (e.g. Creutz and Lagus (2002), Schmid et al. (2004), Sirts and Goldwater (2013), Cotterell et al. (2016b)). Compounds constitute another interesting field of research: being constructed from two or more words, compounds allow to create novel words through the composition of existing ones. Research on compounds addresses the semantic level, for example the relation between the compound's constituents (Ó Séaghdha and Copestake, 2008) or compositionality (Reddy et al., 2011), but also strategies on how to handle compounds in NLP applications like machine translation (e.g. Koehn and Knight (2003) and Cap et al. (2014)), where infrequent and novel compounds pose a challenge due to being insufficiently covered in the training data.

In the context of large language models, there is not much work on the morphological capabilities with regard to complex words. However, the related issue of making words less complex is addressed at the training stage: LLMs do not operate on the word level, but typically on subword pieces such as WordPiece or BPE (Schuster and Nakajima (2012), Sennrich et al. (2016)). There is a large body of research concerning the representation of the training data, mostly studying monolingual settings; and there is a general consensus that frequency-based segmentation approaches are not optimal for morphologically rich languages. For example, Klein and Tsarfaty (2020) argue that a linear splitting into sub-words does not fully capture the morphological complexity of words; similarly, Hofmann et al. (2021) show that a linguistically grounded segmentation can improve a model's performance. Indeed, there are many variants of language-specific PLMs trained on representations that cater to that language's specific demands (e.g. Antoun et al. (2020); Nzeyimana and Niyongabo Rubungo (2022) explicitly model morphological compositionality of linguistically segmented sub words). Jabbar (2023) proposes a linguistically-informed representation for LLM training that is partly based on a database of derivational and inflectional morphology. They do not rely on linear segmentation into concatenable pieces, but instead adapt a representation that can revert to lemmatized forms, making the detokenizing step less straighforward as the pieces cannot just be concatenated, but have to be reconstructed into inflected forms. While our work does not specifically focus on the underlying subword segmentation, we take the general agreement that segmentation strategies are relevant to the model's performance as a motivation to study the understanding of complex words, which to the best of our knowledge has not received much attention so far.

There are relatively few studies that systematically assess the morphological capabilities of PLMs. Shmidman et al. (2023) investigate whether PLMs can distinguish between Hebrew homograph analyses. Hebrew exhibits several types of ambiguities, the resolution of which is intertwined with the word segmentation. Their task corresponds to a word sense disambiguation problem in which they investigate to what extent contextualized embeddings can disambiguate homographs. Weissweiler et al. (2023) analyze the morphological capability of ChatGPT by means of the *Wug*-test (Berko, 1958), in which inflected forms of a nonce word are to be created, thus testing the model's (or a person's in the original version of the test) ability to understand and apply the underlying patterns of the examined morphological operation on unseen data. The paper studies four typologically different languages (English, German, Tamil, Turkish) for different phenomena, such as creating past tense or plural forms as in "*This is a wug. Now there are two* __". They find that ChatGPT performs worse than systems trained specifically for morphological tasks, in particular for English. Haley (2020) presented a similar study. While there are some parallels to our work, we also take on different perspectives by not only looking at the ability to generate word forms, but also to understand and interpret given word forms, including the identification of invalid forms.

Blevins et al. (2023) apply structured prompting to linguistic-framed word- and span-level tagging tasks for English, in which the partially labeled input sentence is iteratively re-fed to the model for the next prediction. They observe a strong performance in few-shot settings for the studied tasks; their findings also indicate that the model's knowledge of linguistic structure is more general than the memorization of the task data. Their study is very interesting, as it aims at obtaining linguistic annotation from the model without retraining, and thus shares a central point of interest with our study. In contrast, our work is realized less formally than a sequence tagging task, but is rather designed as a natural-language question-answer setting. Another difference, of course, is our focus on word-internal structures as opposed to word level.

With a view to studying word understanding and conceptualization, Coil and Shwartz (2023) evaluated PLMs on their ability to paraphrase English noun compounds, comparing sets of existing and novel compounds. They showed that GPT3 achieved a near perfect performance for existing compounds, with a somewhat lower performance for the set of new compounds. Furthermore, they found that correct paraphrases were to a large degree copied from the pre-training data. Working with English compounds, they do not have the issue of decomposing the compound, which is one of the central points of our work. Nonetheless, there is some degree of similarity at the level of compound understanding, even though their work focuses on paraphrasing noun compounds, whereas our study rather looks at a grammatical analysis of compounds, which can be considered a first step in a semantic interpretation.

Another area of interest is the extraction of world knowledge, again mostly limited to English where the word-internal structure is largely irrelevant, for example Petroni et al. (2019) or Hanna and Mareček (2021). In particular for the latter, who study BERT's ability to predict the hypernym of a given word, the ability to process complex words would be an important ability for applying their approach to compounding languages.

3. Data Sets and Methodology

This section gives a short overview of the creation of the test sets, the implementation of the experiments and the evaluation process.

Data Our test sets consist of morphologically complex words with different properties, selected from German newspaper data[1] (13.6 M sentences, 245.9 M tokens). As a basis for the extraction step, the corpus was analyzed with a morphological tool Schmid et al. (2004) to obtain the word-internal structures, for example:

Netznutzungsentgelt:		(*grid usage charge*)
Netz<NN>	nutzen<V>ung<SUFF><NN>	Entgelt<NN>
grid<NN>	*use<V>age<SUFF><NN>*	*charge<NN>*

The analysis contains information about compounding and derivation processes and thus facilitates the search for words of the desired structures for the different tasks; it also provides the basis for the evaluation. In the above example, we can derive that the word is a noun-noun-noun compound with the head *Entgeld* (i.e. the right-most noun). Furthermore, we can retrace the derivational process of the contained verb *nutzen* (*to use*), which was nominalized by means of the suffix *-ung*, resulting in the noun *Nutzung* (*usage*).

The sets of selected words then underwent a manual check to eliminate words with incorrect analyses, misspellings or otherwise undesired properties, such as containing proper names. Not being interested in inflectional morphology, all words in the data set are listed in the lemmatized form.[2]

Methodology The experiments were carried out on the model *gpt-3.5-turbo-instruct* using the function *openai.Completion.create* through the *openai API*. For each prompt, we generated one response with a temperature of 0.2. We went for a comparatively low temperature (in contrast to Coil and

[1] https://data.statmt.org/news-crawl/de/news.2011.de.shuffled.deduped.de

[2] The data set can be found in https://github.com/mariondimarco/ComplexWords_dataset

Shwartz (2023) who set the temperature to 1 for a task requiring more creativity than ours), in order to obtain relatively stable results, while still allowing for some imaginativeness with a view to the rare and sometimes creative target words in our data.

The prompts are formulated in German for all experiments, but for a better readability are translated into English when being discussed in the paper.[3] In some prompts, we included a request such as "answer within one sentence" as this seemed to keep the answers more concise.

Evaluation Evaluating the output of generative language models is non-trivial due to the open output space. In an attempt to avoid extensive manual evaluation, we designed the tasks such that the generated answers can be largely evaluated automatically, typically followed by a manual check: the answers tend to adhere to patterns such as "*the lemma is L*", or are answers to *yes/no* questions.

4. Noun Compounds

In this section, we study the model's ability to analyse and generate complex nouns, both with existing and novel compounds. Compounding is a very productive word formation process in German, where several words are concatenated to form a new one. A compound noun has one *head* (the right-most noun) and one or more *modifier*. The head determines the grammatical properties of the compound (such as part-of-speech and the grammatical gender) and it is often a more general instance of the compound: for example, a *chocolate cake* is a type of *cake* (*Schokoladenkuchen* → *Kuchen*), even though this does not hold for all compounds. The words in the modifier position often do not correspond to the lemmatized form, but can undergo a vowel change. Furthermore, one might need to add or remove characters when joining two words ("*transitional elements*")[4], which makes the task of obtaining a *lemmatized* segmentation more challenging, as illustrated in the examples below:

*H**ü**hn**er**futter* → *H**u**hn Futter* (*chicken feed*)
Farbgefühl → *Farb**e** Gefühl* (*colour sense*)

4.1. Prediction of the Head Noun

The ability to identify the head noun is an important part when processing a compound, as the head contains key information about the word. Table 1 shows the results for 3 prompts with varying specificity with regard to grammatical terms (all are preceded by the request to answer in one sentence):

- **P-1** *What is the head noun of W?*
- **P-2** *What is the head noun of the word W?*
- **P-3** *What is the head noun of the compound W?*

Prompt P-1 yields the worst results with a large proportion of the answers proposing the modifier as the head noun. Answers of the category "*definition*" contain an attempt at explaining the compound. This often demonstrates a good understanding of the compound, but makes no sense in the context of the question, as in the example below:

> Das Kopfnomen von **Farbtreue** ist die Fähigkeit, Farben korrekt und unverfälscht wiederzugeben.
> *The head noun of **colour fidelity** is the ability to correctly and accurately reproduce colours.*

The other prompts introduce the domain of grammar by stating that *W* is a word or compound: this seems to "guide" the model, and results in better answers, about 90 % correct predictions for both variants. This result indicates that the LLM has a generally good understanding of noun compounds, even though it is not perfect, and that this knowledge is not only used implicitly, but can also be accessed through an appropriately formulated prompt.

	prompt1	prompt2	prompt3
head found	119	270	269
modifier	96	26	21
definition	82	–	–
unsplit	1	–	4
other	2	4	6

Table 1: Results for identifying the compound head, comparing three different prompts (N=300).

	prompt1	prompt2	prompt3
head found	101	237	264
modifier	133	58	29
definition	60	1	–
unsplit	1	1	2
other	5	3	5

Table 2: Results for identifying the head of novel compounds, comparing three prompts (N=300).

4.2. Novel Compounds

To investigate the possible role of memorization effects, we look at novel compounds. This is inspired by Coil and Shwartz (2023), who contrast paraphrasing established and novel English compounds; they found that "good" paraphrases of existing compounds tend to have some overlap with the training data. For the task of identifying the head noun, textual overlaps such as "*H is the head noun of W*" are not very likely, but the model might

[3]We list the German prompts in the Appendix.

[4]While there are many patterns to determine the use of transitional elements, there remain some inconsistencies that are difficult to capture on the surface level, e.g. *Buch|druck* (*book printing*) vs. *Büch**er**|regal* (*book shelf*).

still benefit from observing an existing compound in proximity to its head.

The creation of novel compounds is based on random combinations of nouns N1 and N2, for which we checked that their concatenations N1-N2 and N2-N1 were not observed as compounds in the underlying corpus[5] (cf. section 3). Then, one of the pairs N1-N2 or N2-N1 was randomly selected as a novel compound.

A proposed novel compound that is unobserved in our corpus is not necessarily novel with regard to the LM training data; in fact we do not know for sure whether the created compounds are novel to the system. However, looking at the generated words, we are confident that a large majority is unlikely to have occurred before. It is also important to note that the generated compounds are not required to have a reasonable meaning, as the task at hand is entirely grammatical. Some examples are *Wetterschokolade* (*weather chocolate*), *Phonetiktoaster* (*phonetics toaster*), *Realitätsefeu* (*reality ivy*) and *Abwesenheitswichtel* (*absence gnome*).

The general outcome (cf. table 2) is similar to that of the existing compounds, even though slightly worse. Again, we see a considerable number of answers of the category "*definition*" for prompt P-1, even though fewer instances than for the existing compounds. We assume that providing a sort of definition is triggered as a standard reaction to prompt P-1, which is more difficult to answer for novel compounds. The answers often provide a reasonable explanation, but make no sense considering the prompt asked about the *head noun*:

> Das Kopfnomen von **Simulatorzucker** ist ein virtueller **Zucker**, der in **Simulationen** verwendet wird.
> *The head noun of **simulator sugar** is a virtual **sugar** that is used in simulations.*

Overall, these results suggest that the model is able to handle (presumably) unseen compounds, and thus that memorization is not likely to be a relevant factor in this task.

4.3. Analyzing Longer Noun Compounds

For this segmentation task, we look at noun compounds containing 2, 3 and 4 nouns (50 of each) and evaluate the model's ability to derive the length of the compound[6] and to name the individual components, see below for the prompt:

> *Answer in one sentence: How many nouns does the word W contain and what are they (lemmatized)?*

[5]Checking the existence of both concatenations is to ensure that a potential novel compound N1-N2 does not have a "sibling" N2-N1 that could lead to a bias.

[6]Lexicalized words such as *Regenbogen* (*rainbow*) are typically left as one word; there are cases where both splitting and not splitting can be considered correct.

exact match	90
undersplit$_{reasonable}$	39
undersplit$_{invalid}$	2
overlapping results	12
lemma mismatch	5
other	2

Table 3: Segmenting noun-noun compounds into lemmatized nouns (N=150).

Most answers are of the structure "*The word W consists of K nouns: W1, W2, ...*", with *K* being the number of actually listed words in all but 9 cases. For 60 % of the proposed analyses, the segmentation matched exactly with the gold analysis (cf. table 3). For a further 23 % at least one word remained unsplit with an otherwise correct analysis; we consider these analyses as reasonable if the unsplit word is an established noun (such as *Notarzt* (*emergency physician*), or the compound's construction hierarchy is not violated (as in [*Glasboden*] *Boot*) vs. *Glas* [*Bodenboot*] ([*glass floor*] *boat* vs. *glass* [*floor boat*]). In the category *overlapping results* parts of the compound occur several times in the analysis, as in

Target word	Not|arzt|wagen|besatzung *emergency physician vehicle crew*
Answer	Notarztwagen, Besatzung, Notarzt

which suggests that the segmentation task is not trivial. Finally, the category *lemma mismatch* refers to those answers where at least one (otherwise correctly split) noun was not given in the correct lemmatized form, such as *Frauen|fußball* (*women's soccer*) → *Frau**en** Fußball* instead of *Frau Fußball*.

An interesting factor here could also be the subword splitting in the pre-training data: while the word pieces are generally quite short at ca. 3-4 characters, and thus do not even approximate a segmentation into actual noun components, it might be relevant whether the splits occur at noun morpheme boundaries. While splits that do not correspond to the word boundaries do not prevent a word from being segmented correctly, this might still make the task harder, as illustrated in the example:

Target word:	Baby|puder|duft (*baby powder scent*)
Answer:	*Babypuder, Duft*
Subwords:	B, ab, yp, uder, du, ft

5. Derivation

Derivation describes the process of creating new words from an existing one, often through the addition of prefixes and suffixes. There are many productive derivation patterns in German, for example to transform verbs into adjectives or nouns. We evaluate to what extent the model can retrace the

	Prompt1	Prompt2	Prompt3
correct lemma	334	143	297
stem	7	86	8
incomplete	3	2	4
bar-word	4	1	1
noun	–	115	40
other	2	3	–

Table 4: Common verb stem task. (N=350)

underlying stems and understand the meaning of the constructed forms.

5.1. Common Verb Stem

Morphological complexity increases the vocabulary size and data sparsity , which can negatively affect the generalization abilities of a system, as statistics of related words are considered separately. Thus, the ability to derive that two words are related in the sense that they share a common element is important. Because the components are often modified, for example by inserting or deleting characters, as well as adding derivational morphemes, this is not trivial. We test the ability to find the lemma of a verb stem that is "hidden" in a derivation with one of the nominalization suffixes *-ung/-er* or the adjectivization suffix *-bar* in two complex compound nouns. In the example below, the task thus consists in identifying the relevant shared string (the verb stem *beantwort-*), and to output its lemma.

The word part with the target verb stem can occur at the word initial position, in the middle position or in the head position (as in the example below) of the compound.

Beschwerdebrief**beantwort**er, Anfragen**beantwort**ung
***answer**er to complaint letters,* ***answer**ing to requests*
Shared verb: beantworten (*to answer*)

The test set consists of 350 word pairs, of which 334 contain the suffixes *-ung* vs. *-er* and 16 use the suffix *-bar* vs. *-ung/-er*[7]. We compare the following three prompts:

- **P-1** What verb stem occurs in both words A and B? List the lemma.
- **P-2** What common word stem occurs in both words A and B? List the lemma.
- **P-3** What common word stem occurs in both words A and B? List the lemma and the part-of-speech.

Table 4 shows the results: we distinguish between *correct lemma*, *stem* where only the stem is given (e.g. *verfilm* instead of *verfilmen* (*to film*)) and *incomplete* matches such as *geben* instead of *vergeben* (*to give vs to forgive*). In some cases, the proposed answer was either the *bar-word* or one of the *noun* variants. For prompt P-1, the most specific prompt in the sense that it asks for a *verb*, the model solves the task quite well, and outputs the correct lemma for 95 % of the words. This is different for prompt P-2, where the proposed stem is often either a related noun (in most cases one of the input nouns), or the verb stem, which in most cases is not a valid word form by itself. Prompt P-3 sees more answers containing the actual lemma, even though fewer than P-1; it seems that the explicit mentioning of the part-of-speech in the prompt triggered responses that (i) contain an actual valid lemma (as opposed to just stems) in combination with the corresponding part-of-speech, and that (ii) contain verbs rather than nouns as the answer, possibly due to having to decide explicitly on the part-of-speech.[8]

However, even though the task of retrieving the *lemma* of the shared word stem is not completely answered when returning either the stem or a related noun, it still shows that the model identified the relevant word in the broader sense. Thus, the results indicate that the model is capable of establishing a relation between two words sharing a common word stem; for this task, the exact realization of the link (i.e. verb/noun lemma or stem) is of little importance. If one is interested, however, in the actual verb lemma, the results suggest that a precise formulation of the task is important.

5.2. Identifying Invalid Forms

In this experiment, we evaluate the model's knowledge of derivational rules, and in particular the ability to identify ill-formed words: We present the model with artificial words that contain invalid combinations of prefixes and suffixes, but that could pass, at a first glance to a person with only rudimentary knowledge of German, as valid words.

The test set is based on *-bar*-adjectives, which are formed by adding the suffix *-bar* to a transitive verb, analogous to the English *-able*: *fold – foldable* (*falten – faltbar*). Table 5 shows two sets of words: *Set_intrans* contains intransitive verbs with added *-bar*, i.e. incorrect derivations, to test the model's knowledge of what words are appropriate for *-bar* adjectivization. *Set_main* contains valid *-bar*-adjectives.

As a first variant, we nominalize the adjective with the suffix *-keit* (*Knetbarkeit: knetbar+keit* (*kneadability: kneadable+ity*), resulting in valid words. For a simple contrastive set, we just switch the suffixes *-bar* and *-keit* (e.g. **Knetkeitbar*). This is a suffix combination that never exists. To create more subtly incorrect forms, we stack on common prefixes

[7]The comparatively small amount of *-bar*-words is due to the testset containing only nouns and no adjectives.

[8]As we are mainly interested in the correct lemma, the results in table 4 only reflect the answers with regard to the proposed word form, but not the POS.

Set_intrans	krähbar, schlafbar, bleibbar, lachbar, helfbar, springbar, schweigbar, fluchbar, schluchzbar, sausbar, blühbar, leuchtbar, flackerbar, jammerbar, schwelgbar, freubar, kommbar, dauerbar, arbeitbar, hustbar
Set_main	siebbar, baubar, trennbar, backbar, steuerbar, werfbar, druckbar, drehbar, wählbar, waschbar, montierbar, dehnbar, messbar, verstärkbar, stauchbar, teilbar, regulierbar, befüllbar, versenkbar, schmelzbar, zählbar, kühlbar, änderbar, klappbar, faltbar, ausdruckbar, absperrbar, vererbbar, knetbar, entwirrbar

Table 5: Words used as the basis for the "non-word" experiment. *Set_intrans*: incorrect *-bar*-adjectives based on intransitive verbs; *Set_main*: valid and existing *-bar*-adjectives.

Set	**"yes"**	**"no"**
*Set_intrans	19	1
Set_main	30	–
Set_main + *-keit*	30	–
*Set_main_contrastive + *-keit*	5 / 1* / 7**	15 / 2*
*Set_main + *ge-*	30	–
*Set_main + *unge-*	30	–
*Set_main + *unge-...-lich*	14	16
*Set_main + *unge-...-lichkeit*	2 / 27*	1*

Table 6: Answers to the question "*is W a word?*", ignoring the explanation part of the question. A * denotes sets with invalid words.

and suffixes: (i) the prefix *ge-* which is often used to form a past participle; (ii) the negation prefix *un-*; (iii) the adjective suffix *-lich*; and (iv) the nominalization suffix *-keit*. While the created words are meaningless and increasingly absurd (**Ungeknetbarlichkeit*), they are constructed seemingly correctly with regard to the position and order of the added suffixes; furthermore, the respective combinations of *un+ge* and *lich+keit* do exist, for example *ungesehen* (*unseen*) or *Freundlichkeit* (*friendliness*) though not all of them in combination with *-bar*.
We use the following prompt:

> Answer in one sentence: Does the word W exist, and if so, what does it mean?

Table 6 shows the results for the different sets. Interestingly, for the set of incorrect *bar*-adjectives based on intransitive verbs (*Set_intrans*), the system took them for existing words and also provided an explanation that, in the most cases, transports the general meaning of *-bar* in combination with the verb, as in the example below:

**hustbar → ... dass etwas ... gehustet werden kann*
**coughable → ... that something ... can be coughed*

For the set of *-bar*-adjectives based on transitive verbs (*Set_main*), as expected, the LLM recognizes all words as existing and provides a reasonable definition for all but 2 words. This indicates that the LLM is able to understand the meaning of *-bar* added to a verb, but also that it fails to recognize inappropriate contexts. This task requires implicit knowledge about the verb's subcategorization frame (transitive vs. intransitive), as well as understanding of when a rule can be applied. Thus, even though LLMs have been found to have syntactic awareness, this is a challenging task.

The words with *-keit*, i.e. correct nominalization, in *Set_main_keit* are all recognized as such. For the variation with the switched suffixes, about half of the forms were correctly recognized as non-existent. Interestingly, in some cases, the LLM changed the input word to a different, but still false word in the answer (e.g. *Siebheitbar* instead of *Siebkeitbar*); listed with $*$ in table 6. For another 7 words (marked with a $**$), the system used an actual existing word in the answer. This result suggests that the system is generally able to identify blatantly incorrect forms, even though far from perfect.

The 4 bottom rows of table 6 list the results for increasingly complex words. For the sets with the added prefixes *ge-* and *unge-*, the LLM declares to know the words in all cases; the explanation typically refers to the base verb, either through the verb itself or a related noun, and to a lesser extent, through synonyms. For the sets with *un-*, the negation typically figures in the explanation as well. Adding the suffix *-lich* leads to a drop in `yes` answers, but still half of the words are assumed to be correct. Interestingly, the final addition of the noun suffix *-keit* leads to a rise in incorrect `yes` answers. It is notable, however, that in 27 cases, the answer does not cite the original word, but a (still incorrect) variant without the *-lich* suffix, for example:

**Ungeklappbarlichkeit* **un-ge-collapsible-ly-ness*
Ja, das Wort "Ungeklappbarkeit" existiert und bedeutet, dass etwas nicht zusammengeklappt werden kann.
Yes, the word "ungecollapsibleness" exists and means that something cannot be collapsed.

Only two original words are accepted as correct.

These results show that the model does have problems in recognizing invalid forms, even when they are obviously not well-formed. It seems that the extra morphemes are just ignored while those with the most relevant meaning, namely the negation and *-bar*, are interpreted correctly. It might also be the case that the model does not expect incorrectly formed words other than misspellings, typos and possibly typical errors made by language learners, which all are different than the words presented in this experiment. However, the fact that the words in the answers are sometimes "corrected" to slightly different forms suggests that there is a certain insecurity when processing the word.[9]

[9]A similar behaviour was also observed by Weiss-

dim → word	
correct form	235
correct (infl.)	3
head wrong	24
fuge wrong	13
incomplete	22
synonym	2
same word	1

word → dim	
correct form	213
correct (alt.)	2
head wrong	18
fuge wrong	3
incomplete	17
wrong word	7
mod_dim	17
both_dim	10
other	13

Table 7: Creating the word without diminutive form (left); creating the diminutive form (right). (N=300)

5.3. Diminutive Forms

In the previous experiments, the non-concatenative aspects of derivational morphology only played a small role, as they were largely regular, such as the removal of the verb infinitive suffix *-en* to be replaced with *-bar*. Creating diminutive forms involves a wider and less predictable range of non-concatenative operations: in addition to attaching the suffix *-chen* to the head noun, there can be a vowel change, as is often also the case in plural forms or nouns in the modifier position of a compound. Furthermore, the final characters of the word might be removed.

We selected 300 diminutive forms consisting of two or more nouns: more complex forms are likely less frequent, and thus reduce the chance of memorized knowledge. We look at both directions: (i) given a word in diminutive form, find the form without diminutive ending, and (ii) generate the diminutive form for a given word, as in the example below:

(i) *Grundschulst**ü**hlchen* → *Grundschulst**u**hl* (*elementary-school-chair*)

(ii) *Schönwetterw**o**lke* → *Schönwetterw**ö**lk**chen*** (*lovely-weather-cloud*)

weiler et al. (2023), where the generated inflectional variants of a nonce word were mapped to an existing, orthographically similar word (for example *fried* as the past tense of *fride*). They denote this observation as *real-word bias* and formulate the hypothesis that ChatGPT does *not* apply morphological rules, but rather determines the point in the representational space for the answer to be generated given the prompt. If that point is close enough to a real word sharing some properties with the correct answer, namely the required morphological attribute and superficial orthographic similarities, the model rather selects this existing form. To some extent, this idea can also be applied to our scenario, where the model cannot properly fit an incorrect word into its representational space and thus resorts to a slightly less incorrect word, "knowing" that the morphemes employed in the task are highly productive and can thus occur in new words and that they have to adhere to local positional constraints such as *-barkeit* vs. **-keitbar*, but without global understanding of the word's structure.

Table 7 shows the results for the two tasks: while for a 79 % (dim→word) and 71 % (word→dim) of words the correct form could be generated,[10] there is also a number of problems to be observed. The category *head wrong* refers to words where the head noun is ill-formed, for example with the wrong *Umlaut* and/or missing characters. Similarly, *fuge wrong* means that the transitional element between the words is incorrect. In some cases, one of the two (or more) nouns of the compound was lost (*incomplete*). Also, in some cases words were replaced with either synonyms or different (but somewhat similar orthographically) words (*wrong word*). In the second task, we could also observe word forms with the diminutive suffix to the modifier or even both nouns, as shown in the examples below:

Ausflugsort → **Ausfl**ü**g**chen**ort*
(*outing-destination*: *nice place for outings*)

Tannenwald → **T**ä**nn**chen**w**ä**ld**chen*** (*pine forest*)

While the errors of the types *head/fuge wrong* suggest that the non-concatenative operations pose a challenge, the other error types show that there are also problems at the semantic level (i.e. missing parts, reverting to different words), as well as concerning the question of where to apply the morphological rule. In particular the double application of the diminutive suffix is reminiscent of the issues observed in the previous experiment.

6. Word-Internal Negation

Negation has been shown to be tricky in many NLP tasks (for example Kassner and Schütze (2020) in masked LMs). Negation can be realized syntactically as well as morphologically through the addition of prefixes such as *un-* (*un-*).

Compounding is very productive in German, and there exist words with internal morphological negation, for example *senioren<u>un</u>freundlich* (*unfriendly to elderly people*). In the following, we examine the model's ability to process such words by asking for the non-negated form of the words: This task requires the model to first identify how and where in the word the negation is realized (*-un-*), and then to output the same word without negation. It is noteworthy, though, that the set of negated adjectives and nouns in the head position is restricted to words that lend themselves to negation, thus making the task somewhat easier.

We selected 81 nouns of the structure `N -un-` $\texttt{N}_{keit/heit}$, where the second noun is a nominalized adjective (with the suffix *-keit/heit*), and 137 adjectives of the structure `N -un- ADJ`.[11]

[10] *correct(infl)*, *correct(alt)*: the inflection differs from the lemma, or an alternative diminutive suffix was used.

[11] For both sets, the respective positive forms exist in the corpus. While most of the words can have a positive

	ADJ	NOUN
positive form found	118	75
synonym	13	3
same word	3	1
incomplete	2	1
other	1	1

Table 8: Generating the non-negated form.

We used the following prompt:

What is the non-negated form of the word W?

Table 8 shows the results of this task: for both sets, the model was able to generate the non-negated form for a majority of the words. The main error type observed in both categories is the outputting of a synonym of the non-negated word, for example

*konjunktur**un**empfindlich* → *konjunktursensibel*
unaffected by / sensitive to economic fluctuations

While this is, strictly speaking, not a correct answer to the question, where we expected the non-negated form of the exact same word, it still demonstrates understanding of the word meaning, but it can rather be considered as the answer to the question of the opposite of the word. Only in very few cases, the answer is wrong, by outputting either the same word, or an incomplete or otherwise ill-formed word. In one case, the model combined the removal of the negation morpheme with the opposite of the adjective, i.e. a double removal of negation, resulting in a non-existing word:

*schmutz**un**anfällig* → **sauberanfällig*
***un**susceptible to dirt* → *susceptible to clean*.

In summary, the system demonstrated a solid understanding of negation, even though the question is not always answered technically correct by proposing a synonym with the opposite polarity. However, this might be a more natural way to handle this question in natural language, than to give the same word without the negation morpheme.

7. Discussion

Our experiments demonstrate that the LLM has in general a good understanding of complex words, even though there are some interesting errors to be observed. We focus mostly on grammatical tasks, in part for the sake of a straightforward evaluation, but also to assess how accessible this type of information is. In fact, our experiments suggest that we can obtain morphological-structural information (such as the head of a compound) or particular variants of a word form (such a non-negated form), but the outcome is very dependent on the prompt formulation. Furthermore, we observed a tendency to revert to synonyms instead of the original word in an answer. While this is a good strategy when explaining an (unknown) term, and might have been supported in the pre-training, this is a somewhat undesired property for morphological analysis, where we are interested in the very same word.

We initially asked the question whether the LLM's generally assumed syntactic-grammatical comprehension at sentence level is also established at the word-level. This leads to the question whether the model actually understands and applies morphological rules for derivation, or whether the tasks are solved based on memorization or other shortcuts. Here, the interpretation is rather unclear: the results for the novel compounds show that the model can solve this task for unseen words as well, thus suggesting that memorization effects do not play a role. On the other hand, the task of identifying ill-formed words clearly showed that the model has major difficulties, and thus lacks knowledge of the respective morphological rules. Summarizing, it is not trivial to answer this question, in particular when assuming that the task of recognizing ill-formed words is more difficult than identifying a compound's head noun and possibly not a task that is natural to the model's overall pre-training objective.

Finally, there is the question of subword-segmentation: Most words are broken into subwords that for the most part do not correspond to linguistically meaningful units. In contrast to the sentence-level, where the tokenization into words can typically be assumed as correct (i.e. through the white-spaces in the text, at least for languages like English), the segmentation at the subword level is often linguistically sub-optimal. In particular, some words are segmented such that the individual nouns cannot be obtained (cf. example in section 4.3). While there is a lot of evidence that non-linguistically informed segmentation methods are not optimal, the effect of the quality of subword splitting on the understanding of complex words remains still unclear and makes an interesting topic for future work.

8. Conclusion

We designed several tasks and data sets to evaluate an LLM's ability to understand morphologically complex words; an area that has not received much interest so far, even though the studied properties are important for processing morphologically rich languages. We showed by means of querying for components of a complex word that the model has a good understanding of complex words, but fails at recognizing invalid forms. However, it remained unclear to what extent the model actually understood and applied derivational rules, and to what extent its abilities are based on other effects.

form, this does not make much sense for some words, for example *Wahlunregelmäßigkeit* (*election irregularity*).

9. Limitations

An obvious limitation is that we only investigated German and the relatively small size of our semi-manually designed data sets. Extending this approach to other languages, including low-resource languages, to cover a wider range of morphological phenomena, constitutes an interesting next step in understanding how a model like GPT processes and interprets complex words.

Similarly, we only look at the understanding of complex words from a monolingual perspective, even though most LLMs are heavily multilingual. Exploring relations between complex words across languages might provide further insights.

Finally, our interest in this study is mainly focused on a mostly superficial grammatical level such as identifying the head noun of a compound, and leaves out in large parts the finer points of a semantical interpretation of morphologically complex words, such as examining the relations between the components of a compound, for example *Apfelsaft* (*apple juice*: made from apples) vs. *Hustensaft* (*cough juice*: to remedy coughing).

10. Acknowledgements

The work presented in this paper was carried out while both authors were at the Center for Information and Language Processing, LMU Munich.

This publication was supported by LMUexcellent, funded by the Federal Ministry of Education and Research (BMBF) and the Free State of Bavaria under the Excellence Strategy of the Federal Government and the Länder; and by the German Research Foundation (DFG; grant FR 2829/4-1).

11. Bibliographical References

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based Model for Arabic Language Understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Jean Berko. 1958. The Child's Learning of English Morphology. *WORD*, 14:150–177.

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.

Fabienne Cap, Alexander Fraser, Marion Weller, and Aoife Cahill. 2014. How to Produce Unseen Teddy Bears: Improved Morphological Processing of Compounds in SMT. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 579–587, Gothenburg, Sweden. Association for Computational Linguistics.

Albert Coil and Vered Shwartz. 2023. From Chocolate Bunny to Chocolate Crocodile: Do Language Models Understand Noun Compounds? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2698–2710, Toronto, Canada. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016a. The SIGMORPHON 2016 Shared Task—Morphological Reinflection. In *Proceedings of the 2016 Meeting of SIGMORPHON*, Berlin, Germany. Association for Computational Linguistics.

Ryan Cotterell, Arun Kumar, and Hinrich Schütze. 2016b. Morphological Segmentation Inside-Out. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2325–2330, Austin, Texas. Association for Computational Linguistics.

Mathias Creutz and Krista Lagus. 2002. Unsupervised Discovery of Morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, pages 21–30. Association for Computational Linguistics.

Allyson Ettinger. 2020. What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Coleman Haley. 2020. This is a BERT. Now there are several of them. Can they generalize to novel words? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 333–341, Online. Association for Computational Linguistics.

Michael Hanna and David Mareček. 2021. Analyzing BERT's Knowledge of Hypernymy via Prompting. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 275–282, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Haris Jabbar. 2023. MorphPiece : Moving away from Statistical Language Representation. ArXiv:2307.07262.

Nora Kassner and Hinrich Schütze. 2020. Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Stav Klein and Reut Tsarfaty. 2020. Getting the ##life out of living: How Adequate Are Word-Pieces for Modelling Complex Morphology? In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 204–209, Online. Association for Computational Linguistics.

Philipp Koehn and Kevin Knight. 2003. Empirical Methods for Compound Splitting. In *10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary. Association for Computational Linguistics.

Karim Lasri, Tiago Pimentel, Alessandro Lenci, Thierry Poibeau, and Ryan Cotterell. 2022. Probing for the Usage of Grammatical Number. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8818–8831, Dublin, Ireland. Association for Computational Linguistics.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic Knowledge and Transferability of Contextual Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Antoine Nzeyimana and Andre Niyongabo Rubungo. 2022. KinyaBERT: a Morphology-aware Kinyarwanda Language Model. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5347–5363, Dublin, Ireland. Association for Computational Linguistics.

Diarmuid Ó Séaghdha and Ann Copestake. 2008. Semantic Classification with Distributional Kernels. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 649–656, Manchester, UK. Coling 2008 Organizing Committee.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An Empirical Study on Compositionality in Compound Nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

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).

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and Korean Voice Search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152.

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.

Naomi Shapiro, Amandalynne Paullada, and Shane Steinert-Threlkeld. 2021. A Multilabel Approach to Morphosyntactic Probing. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4486–4524, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Avi Shmidman, Cheyn Shmuel Shmidman, Dan Bareket, Moshe Koppel, and Reut Tsarfaty. 2023. Do Pretrained Contextual Language Models Distinguish between Hebrew Homograph Analyses? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 849–864, Dubrovnik, Croatia. Association for Computational Linguistics.

Kairit Sirts and Sharon Goldwater. 2013. Minimally-Supervised Morphological Segmentation using Adaptor Grammars. *Transactions of the Association for Computational Linguistics*, 1:255–266.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

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.

12. Language Resource References

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).

A. German Prompt Formulations

We used German prompts for all experiments, but gave the English translation in the main part of the paper for better readability. Below, we list the used prompt formulations.

Identification of the head noun in section 4.1

- **Prompt-1** *Antworte in einem Satz: Was ist das Kopfnomen von W?*
- **Prompt-2** *Antworte in einem Satz: Was ist das Kopfnomen des Wortes W?*
- **Prompt-3** *Antworte in einem Satz: Was ist das Kopfnomen des Kompositums W?*

Analyzing longer compounds in section 4.3

- **Prompt** *Antworte in einem Satz: Aus wievielen Nomen besteht das Wort W und wie lauten sie (lemmatisiert)?*

Common verb stem in section 5.1

- **Prompt-1** *Welcher Verbstamm kommt in den beiden Wörtern W1 und W2 vor? Nenne das Lemma.*
- **Prompt-2** *Welcher gemeinsame Wortstamm kommt in den Wörtern W1 und W2 vor? Nenne das Lemma.*
- **Prompt-3** *Welcher gemeinsame Wortstamm kommt in den Wörtern W1 und W2 vor? Nenne das Lemma und die Wortart.*

Identifying invalid words in section 5.2

- **Prompt** *Antworte in einem Satz: Gibt es das Wort W und wenn ja, was bedeutet es?*

Diminutive Form Tasks in section 5.3

- **Prompt-1** *Wie lautet das Wort W ohne Diminutivendung?*
- **Prompt-2** *Wie lautet das Wort W im Diminutiv?*

Negated Words in section 6

- **Prompt** *Wie lautet die nicht negierte Form des Wortes W?*