Unlike “Likely”, “Unlike” is Unlikely: BPE-based Segmentation hurts Morphological Derivations in LLMs

Paul Lerner and **François Yvon**
Sorbonne Université, CNRS, ISIR
75005, Paris, France
lerner@isir.upmc.fr, yvon@isir.upmc.fr

Abstract

Large Language Models (LLMs) rely on subword vocabularies to process and generate text. However, because subwords are marked as initial- or intra-word, we find that LLMs perform poorly at handling some types of affixations, which hinders their ability to generate novel (unobserved) word forms. The largest models trained on enough data can mitigate this tendency because their initial- and intra-word embeddings are aligned; in-context learning also helps when all examples are selected in a consistent way; but only morphological segmentation can achieve a near-perfect accuracy.

1 Introduction

Large Language Models (LLMs) constitute a workhorse of modern Natural Language Processing applications, owing to their unprecedented ability to generate syntactically correct, semantically coherent, and pragmatically relevant utterances, responses to a wide array of queries, in a growing number of languages. As recent studies have shown, during their training process, LLMs also acquire some sort of morphological abilities, e.g., to generate inflected forms for known and unknown lemmas (Weissweiler et al., 2023) – at least when they follow regular morphological patterns (see also Hofmann et al., 2020; Mortensen et al., 2024). These abilities extend even to previously unknown languages, given that some examples of the targeted patterns are provided in the prompt (Tanzer et al., 2024; Zhang et al., 2024). Such morphological knowledge is essential to achieve good performance in constrained (e.g., Machine Translation) as well as unconstrained text generation applications. The ability to manipulate and recombine substrings and to handle unknown word forms can be attributed to the use of subword vocabularies, e.g., relying on Byte Pair Encoding (BPE; Gage, 1994; Sennrich et al., 2016).

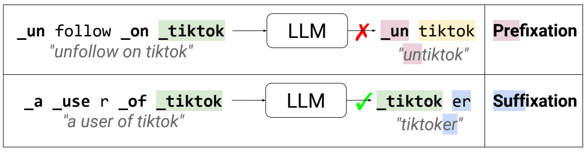

Figure 1: In BPE tokenization, marking word-initial tokens with “_” hinders the generation of prefixed forms (e.g., “`_un tiktok`”), as they do not share any token with their base (e.g., “`_tiktok`”). Identical tokens are highlighted in the same color.

In this contribution, we show that if BPE-based tokenizers enable morphological generalization, they do not handle all morphological processes equally well. The reason, we claim, is that BPE marks word-initial substrings with a special character “_”, to make tokenization reversible (Kudo and Richardson, 2018).[1] Therefore, suffixed and prefixed forms are handled differently: once tokenized, suffixed forms such as “*tiktoker*” may share a subword with their base “*tiktok*”, implying also some semantic similarity. Crucially, this cannot happen with prefixed forms like “*untiktok*”, which, even assuming a morphologically plausible tokenization “`_un tiktok`”, are represented using a distinct token “`tiktok`”, unrelated to the base “`_tiktok`” (see Figure 1; Hofmann et al., 2020).[2]

We present here experiments that highlight this problem in a very controlled setting. For this, we

[1]Equivalently, Sennrich et al. (2016) marked intra-words with “@@”, e.g., “`tiktok @@er`”. Marking the end of words instead of their start would hinder suffixations. The issue is identical for all subword tokenizers that we are aware of: BPE, Unigram (Kudo, 2018), and WordPiece (Wu et al., 2016), as they all mark word-initial substrings to make tokenization reversible. A similar case is made by Hofmann et al. (2021) about WordPiece, as discussed in Section 5. We simply focus on BPE as it has now been widely adopted by all modern LLMs (e.g., GPT-4, (OpenAI, 2023); Llama-3, (Llama Team, 2024); and Gemma, (Gemma Team, 2024)).

[2]Word-internal tokens only make their way to the tokenizer’s vocabulary if supported by enough prefixed forms in the training corpus; otherwise, such derivatives are over-segmented, e.g., “`_un tik t ok`” (Lerner and Yvon, 2025).

Proceedings of the 31st International Conference on Computational Linguistics, pages 5181–5190
January 19–24, 2025.

consider two regular affixation processes in English and French: negative prefixations (e.g., EN "*un-*", FR "*in-*") and adverbial suffixations (e.g., EN "*-ly*", FR "*-ment*"). As both processes apply to adjectives, we can compare on a fair basis the capacities of LLMs to generate prefixations and suffixations of the same set of lexemes. Our experiments include both attested adjectival bases and nonce words. We find, across several LLM families and sizes, that (i) LLMs often fail to derive new words via prefixation compared to suffixation; (ii) the cases where prefixation is successful may be explained by the alignment between word-initial and word-internal embeddings of the same string (e.g., when "**`_tiktok`**" ≈ "`tiktok`" in the embedding space), which is dependent on the model size and amount of pretraining data; (iii) this tendency can be mitigated with in-context learning (ICL), especially with a consistent selection of ICL examples; (iv) the issue disappears when using a morphological segmentation, which leads to near-perfect accuracy, for both prefixations and suffixations.

2 Derivational Morphology

Derivational Morphology is central to the structure of the lexicon, so as to move away from the arbitrariness of the sign (De Saussure, 1916; Lieber, 2010; Corbin, 2012). Affixation is cross-linguistically the most common process that human languages use to derive new lexemes (Štekauer, 2012; Goethem, 2020). For example, Turkish's *-li* attaches to nouns to make personal nouns (e.g., *şehir* 'city' → *şehirli* 'city dweller'); Chinese's *-xué* attaches to nouns to make nouns meaning 'the study of X' (e.g., *dòngwù* 'animal' → *dòngwùxué* 'zoology'); Samoan's *fa'a-* attaches to nouns to make verbs meaning 'make X' (e.g., *goto* 'sink' → *fa'agoto* 'make sink', Lieber, 2010). In this paper, we study English and French as mere examples to motivate our finding about BPE, which formally applies to any text in any language. Regular affixation processes are routinely used to form *neologisms*, new lexemes or terms, either in everyday conversations or in specific domains (Daille, 2017; Cartier et al., 2018; Lerner and Yvon, 2025).

Formally, *prefixation* operates at the beginning of a lexeme (e.g., "*untiktok*"), whereas *suffixation* applies at lexeme' ends (e.g., "*tiktoker*"). This implies, as discussed above, that the two types of derivative will be handled differently by subword tokenizers. Affixation may additionally cause phonological or graphemic change(s), resulting in variation (*allomorphy*) in the surface realization of some lexemes (Lieber, 2010). This is another cause of possible divergence between the tokenization of a base and a derived lexeme. In our experiments, we make sure to only consider cases of purely concatenative affixations,[3] as in the above examples, to isolate the tokenization challenge from other segmentation issues. In English and French, other differences, not developed here, between prefixation and suffixation are that the latter tends to play a more syntactic role (e.g., converting adjectives to adverbs with "*-ly*") while the former holds a more semantic role (e.g., negating adjectives with "*un-*").

Lang.	Affix	Definition
EN	un-	Not <base>
FR	*in-*	*Qui n'est pas <base>*
EN	-ly	In a <base> manner
FR	*-ment*	*D'une manière <base>*

Table 1: Affixations and associated definition templates

3 Methods

3.1 Definition to Word Generation

To measure differences in the way suffixations and prefixations are handled by LLMs, we consider the simple morphological task of generating a lexeme from its definition, framed here as a text-to-text problem (Brown et al., 2020; Raffel et al., 2020), following Lerner and Yvon (2025). Given the definition of a lexeme (e.g., "*a user of tiktok*"), an LLM is prompted to generate the derivative (e.g., "*tiktoker*"), cf. Figure 1. Models are prompted in the same language as the definition (<def>) and the target lexeme, i.e. (i) EN: "`<def> defines the term :`"; (ii) FR: "`<def> définit le terme :`". The expected continuation is the derived form. Definitions alway include the base lexeme and unambiguously correspond to either a prefixed or a suffixed derivative (Table 1).

3.2 In-Context Learning

LLMs can further generalize to such tasks by leveraging In-Context Learning. Our early results suggested that LLMs were not too sensitive to the exact

[3]This means that valid morphological segmentations will always be either "`<prefix> <base>`" for prefixations or "`<base> <suffix>`" for suffixations.

prompt formulation, but mostly leveraged ICL examples, consistently with prior work (e.g., Zebaze et al., 2024). We thus use five ICL examples in each prompt, formatted as above, separated by the three characters ###, which serve as end-of-sequence signal. Here is an example from the ADJ-EN dataset, using a single ICL example: "*Not pluvial defines the term : unpluvial ### Not lightfast defines the term :*" (the model should generate "*unlightfast*").

We limit the number of ICL examples to five to keep a reasonable input length.[4] We compare two ICL selection methods: (i) Random sampling, examples can be either a prefixation or a suffixation; (ii) Morphological: sampling only prefix (resp. suffix) for prefix (resp. suffix) generation tasks.

3.3 Large Multilingual Models

We conduct experiments with three model families: BLOOM (BigScience et al., 2023), CroissantLLM-1.3B (Faysse et al., 2024), and Llama-2-7B (Touvron et al., 2023), including various sizes for BLOOM, ranging from 560M to 7.1B parameters. All models are multilingual and cover EN and FR to different degrees: BLOOM is highly multilingual, trained on 46 natural languages; CroissantLLM is bilingual, trained on an equal share of EN and FR; Llama-2 is mostly trained on EN (89.70%) but does include some FR (0.16%).

3.4 Segmentation

We compare two segmentation strategies:

(i) BPE, used by all studied LLMs. Keeping the same example as above, the base and derived word are tokenized as follow by BPE (for BLOOM but beginning of words are always marked by BPE, regardless of the LLM):

pluvial `_pluv ial`
unpluvial `_un pl uv ial`

Notice how the derived word does not include the tokens of its base.

(ii) Morphological segmentation, where we enforce that derived words in the ICL samples share tokens with their base by adding an extra space to the affix. In that case, the output is expected to be also space separated. For example:

un pluvial `_un _pluv ial`

3.5 Controlled Datasets

We perform controlled experiments, where each base has one derived prefixation and suffixation.

Dataset	Base	Prefixation	Suffixation
ADJ-FR	démontable	indémontable	démontablement
ADJ-EN	lightfast	unlightfast	lightfastly
PSEUDO-FR	géniable	ingéniable	géniablement
PSEUDO-EN	orionful	unorionful	orionfully

Table 2: Examples of a base and its derivatives for each dataset

We study two regular affixations that apply to adjectival bases: (i) negative prefixations (EN: "*un-*", FR: "*in-*"); (ii) adverbial suffixations (EN: "*-ly*", FR: "*-ment*"), paired with the definitions listed in Table 1, e.g.: (i) "*Not lightfast*" → "*unlightfast*"; (ii) "*In a lightfast manner*" → "*lightfastly*".

We experiment with two sets of bases, in each language: (i) ADJ, attested adjectives from MorphyNet (Batsuren et al., 2021), which is built upon Wiktionary; (ii) PSEUDO, pseudo-words generated with UniPseudo (New et al., 2024) (see examples for each dataset in Table 2). As explained above, we restrict ourselves to purely concatenative affixation and avoid allomorphy phenomena using "morphotactic" rules described in Appendix A. MorphyNet has fewer samples in FR than EN, and FR morphotactics are more strict, so ADJ-FR contains 2,313 adjectival bases, i.e. 4,626 derived words (one prefixation and suffixation per base), while ADJ-EN contains 14,455 bases. Pseudo-adjectives are generated with UniPseudo, using a character n-gram model trained with attested adjectives. In each language, we first generate 5,000 nonce words of L letters, for $L \in [\![6, 12]\!]$. After filtering these with morphotactic rules, we obtain PSEUDO-FR (comprising 8,507 bases) and PSEUDO-EN (29,177 bases). The datasets are equally and randomly split in ICL-test splits, without overlap between bases.[5]

4 Results

4.1 Prefixations vs. Suffixations

Figure 2 displays our main results on the four datasets with three different model families: BLOOM, CroissantLLM-1.3B, and Llama-2-7B (detailed scores are in Table 4 in Appendix B). We use Exact Match (EM), also known as *accuracy* to evaluate generation (Cotterell et al., 2016). Clearly, with standard BPE segmentation, suffix generation is overall far superior to prefix generation (e.g., 26.2 EM for prefixes vs. 56.0 for

[4]Early experiments suggest that the difference between prefixes and suffixes is only stronger with fewer examples.

[5]See Appendix C for implementation details and github.com/PaulLerner/neott for code and data.

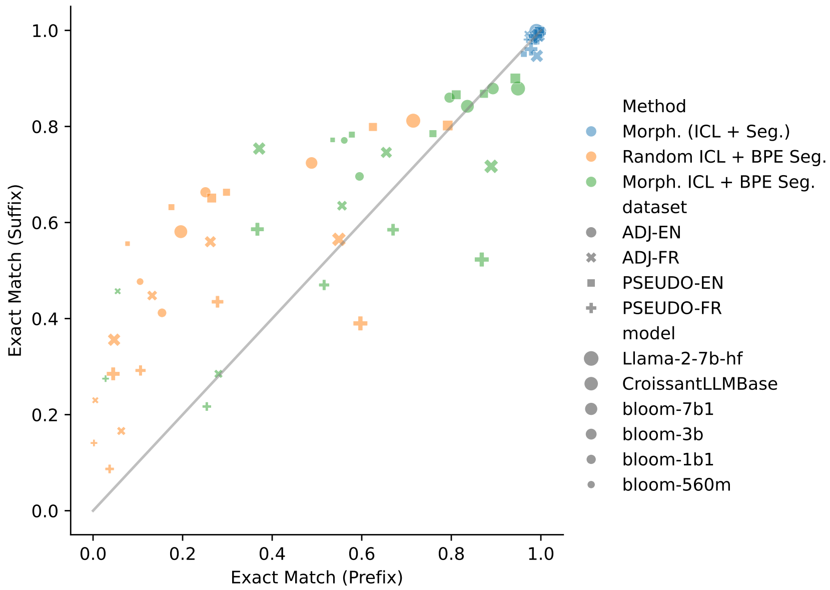

Figure 2: Exact match scores for prefixes vs. suffixes for four datasets (attested adjectival bases and pseudo-words, EN and FR; plotted with different shapes), three model families and four BLOOM model sizes ranging from 560M to 7.1B parameters (plotted with different sizes), according to ICL examples and segmentation method (different colors). Most points are above the line $y = x$, because suffixes are better generated than prefixes.

suffixes, with BLOOM-7.1B on FR attested adjectives; above the $y = x$ line). Errors in prefixations also include cases of morphotactically incorrect forms, with the generated prefixes containing extraneous letters, dashes, or spaces (e.g., "inc*grandiose*", "in-*onirique*", or "in *cognitive*", with BLOOM-7.1B on FR attested adjectives). We also find that prefixes are sensitive to the choice of ICL examples: selecting only prefixes (resp. suffixes) for prefix (resp. suffix) prediction helps to reduce the gap (green vs. orange dots). This finding is consistent with Hofmann et al. (2024) who find that LLM's generalize through analogies rather than rules. We finally observe that Llama outperforms BLOOM and CroissantLLM for this task.

Morphological segmentation, on the other hand, solves the initial- vs. intra-word tokenization issue and yields near-perfect accuracy, for both prefixes and suffixes and all models (blue vs. green dots).[6] Figure 2 shows that even small versions of BLOOM, with 560M or 1.1B parameters (small dots), achieve near-perfect accuracy for prefixes with a morphological segmentation, when the corresponding Exact Match score was close to zero with BPE segmentation. BLOOM-7.1B is still able to correctly generate some prefixed form, probably due to its larger number of parameters. These results are consistent on the four datasets, i.e., for both attested adjectival bases and pseudo-words, in EN and FR.

4.2 Initial- vs. Intra-word Alignment

In this section we ask: how is it possible at all for BPE-based models to generate prefixations? We argue that, like for suffixations where the model simply needs to copy the base (e.g., "**`_tiktok`**") and append a suffix (e.g., "`er`"),[7] for prefixations the model first needs to generate the prefix (e.g., "**`_un`**") then an intra-word token whose representation is close to that of the base (e.g., "`tiktok`"), therefore to model the similarity between the two tokens (e.g., "**`_tiktok`**" ≈ "`tiktok`").

We find that, when a string has dedicated embeddings respectively covering word-initial and word-internal occurrences (e.g. "**`_like`**" and "`like`" or "**`_vraisemblable`**" and "`vraisemblable`"), both are often aligned, i.e., close in the embedding space. To evaluate this, for each pair pairs of vocabulary units of the form $(\_x, x)$ made of a word-initial and a matched word-internal vocabulary entry, we compute the cosine similarity of $\_x$ with all exist-

[6] Note that, while suffixations can always be tokenized by BPE as "`<base> <suffix>`" (e.g., "**`_lightfast`** `ly`"), the optimal tokenization (according to BPE) may not necessarily preserve the base (e.g., "**`_light`** `fastly`"). This explains why morphological segmentation also improves suffixation results.

[7] Empirically, we find across all models and datasets that BPE-based models tend to copy the base tokens at a 63% rate in average when generating suffixations.

Model	# Pairs	# Intra	P@1
CroissantLLM-1.3B	3,771	14,296	71.9
BLOOM-7.1B	13,365*	111,326	76.3
Llama-2-7B	5,272	15,590	83.0

Table 3: Alignment between embeddings of word-initial types and the corresponding word-internal variant, for three models. *BLOOM's vocabulary contains a lot of noise so we evaluate only on fully Latin strings (matching `[A-Za-z]`), otherwise P@1 would drop to 65.4.

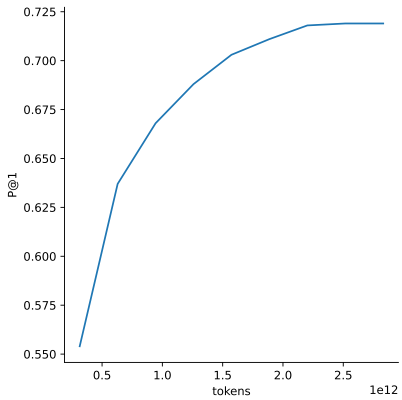

Figure 3: Alignment between embeddings of word-initial types and the corresponding word-internal variant for various checkpoints of CroissantLLM, according to the number of tokens used in training (in trillions).

ing word-internal entries and measure the ability to retrieve the matched entry x with Precision@1 (P@1). Depending on the model, we find P@1 values ranging from 71.9 to 83.0, reported in Table 3. These values are well correlated with the EM scores for prefixes reported above (for the four BPE-based BLOOM models, we find Pearson $r = 0.639, p < 0.01$, across the four datasets).

This finding is consistent with Itzhak and Levy (2022), who find that word embeddings encode the string of characters that compose it; and Tytgat et al. (2024) who find that word embeddings are sensitive to surface similarities (e.g. edit distance).

Figure 3 shows that alignment increases with the number of tokens seen in training: for CroissantLLM, P@1 increases from 55.2 (after 300B tokens) up to 71.9 after 3T (again correlated with EM scores of prefixes of BPE-based models with Pearson $r = 0.338, p < 0.05$, across the four datasets). Therefore, gigantic amounts of data are used to implicitly learn an alignment that could be made explicit using morphological segmentation.

5 Related Work

Our framing of Word Derivation somewhat resembles the *Reverse Dictionary* task (Hill et al., 2016; Pilehvar, 2019). However, Reverse Dictionary is an Information Retrieval task that consists of mapping the representation of a definition to an existing word embedding. On the contrary, we design here a fully *generative* task. Our work is more related to Lerner and Yvon (2025) who leverage definitions to translate neologisms more accurately.

Hofmann et al. (2021) and Truong et al. (2024) are also interested in derivational morphology and LLMs but consider binary classification tasks that assess whether the LLM "understands" words, while we are interested in the actual generation of new forms. Our results are consistent with theirs: notably, Hofmann et al. (2021) also find that LLMs are unable to process prefixations, compared to suffixations, for the same reason. They also find that enforcing morphological segmentation improves performance. Hofmann et al. (2020) is similar to our work but always relies on morphological segmentation, except in their preliminary experiment.

Oh and Schuler (2024) and Pimentel and Meister (2024) discuss another effect of BPE marking the beginning of words: the miscomputation of word probabilities, an indicator of word surprise used in psycholinguistic studies. Both propose a simple rescaling method to recover the correct values.

6 Discussion

BPE is ubiquitous in NLP as virtually all LLMs depend on it. However, marking strings in the beginning of words leads to caveats that are overlooked. We show that this faulty tokenization limits the ability of LLMs to generate prefixations, a morphological process that is however productive in many languages. Such defects in morphological abilities may partly explain the recurrent difficulties of LLMs to generate a sufficiently large number of new lexemes, as attested by low Type-to-Token Ratio scores in generated texts (Muñoz-Ortiz et al., 2024). We also show that an accessible solution is morphological segmentation, which enables even "small" models (of a few hundred million parameters) to reach near-perfect generation accuracy.

Limitations

We study only two languages: English and French. However, we focus on a formal issue of the BPE method, which would be identical for any text and therefore any language. We assume that this caveat would affect only more strongly less-resourced languages.

We are limited to one prefixation and one suffixation per language. This restriction was inevitable to allow stratified data generation (Section 3.5): the chosen negative prefixations and adverbial suffixations are very regular in English and French, both can be applied to any adjective. However, formally, the affixation process is identical regardless of the actual affix, be it *-ly*, *-ation*, or *-ical*.

Hofmann et al. (2021) had already pointed out the issue of marking beginning of words with WordPiece (instead of BPE), and also proposed to fix it by leveraging morphological segmentation. However, we propose a new framework (generation instead of classification) and provide additional analysis to understand the phenomenon through in-context learning (Figure 2), alignment of initial- and intra-word embeddings (Table 3), and amount of pretraining data (Figure 3). Additionally, we conduct extensive experiments on three different LLM families, while Hofmann et al. (2020, 2021) only use BERT (Devlin et al., 2019).

We propose to use morphological segmentation to solve the issue with the BPE tokenizer. This, however, is easier said than done: BPE has the advantage of being language-agnostic and therefore allows transfer learning between languages within a multilingual language model. In contrast, we are not aware of a morphological segmentation method that could be applied to all languages. It would most likely require a language identification pipeline followed by language-specific segmentation.

Acknowledgments

We thank Lichao Zhu and Ziqian Peng for their helpful feedback on the initial draft of this article. We also thank the anonymous reviewers for their knowledgeable comments.

This research was funded by the French "Agence Nationale de la Recherche" (ANR) under the project MaTOS - "ANR-22-CE23-0033-03". This work was performed using HPC resources from GENCI–IDRIS (Grant 2023-AD011014881).

References

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2021. Morphynet: a large multilingual database of derivational and inflectional morphology. In *Proceedings of the 18th sigmorphon workshop on computational research in phonetics, phonology, and morphology*, pages 39–48.

BigScience, Teven Le Scao, and et al. 2023. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv preprint*. ArXiv:2211.05100 [cs].

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Emmanuel Cartier, Jean-François Sablayrolles, Najet Boutmgharine, John Humbley, Massimo Bertocci, Christine Jacquet-Pfau, Natalie Kübler, and Giovanni Tallarico. 2018. Détection automatique, description linguistique et suivi des néologismes en corpus: point d'étape sur les tendances du français contemporain. In *6e Congrès Mondial de Linguistique Française-Université de Mons, Belgique, 9-13 juillet 2018*, volume 46, pages 1–20. EDP Sciences.

Danielle Corbin. 2012. *Morphologie dérivationnelle et structuration du lexique*. Walter de Gruyter. Google-Books-ID: AYwjAAAAQBAJ.

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

Béatrice Daille. 2017. *Term Variation in Specialised Corpora: Characterisation, automatic discovery and applications*, volume 19 of *Terminology and Lexicography Research and Practice*. John Benjamins Publishing Company, Amsterdam.

Ferdinand De Saussure. 1916. *Cours de linguistique générale*, volume 1. Otto Harrassowitz Verlag (1989 reedition).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Manuel Faysse, Patrick Fernandes, Nuno Guerreiro, António Loison, Duarte Alves, Caio Corro, Nicolas Boizard, João Alves, Ricardo Rei, Pedro Martins, Antoni Bigata Casademunt, François Yvon, André Martins, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. CroissantLLM: A Truly Bilingual French-English Language Model. *arXiv preprint*. ArXiv:2402.00786 [cs].

Philip Gage. 1994. A New Algorithm for Data Compression. *Computer Users Journal*, 12(2):23–38. Place: USA Publisher: R & D Publications, Inc.

Google Gemma Team. 2024. Gemma 2: Improving Open Language Models at a Practical Size. *arXiv preprint*. ArXiv:2408.00118.

Kristel Van Goethem. 2020. Affixation in morphology.

Felix Hill, Kyunghyun Cho, Anna Korhonen, and Yoshua Bengio. 2016. Learning to Understand Phrases by Embedding the Dictionary. *Transactions of the Association for Computational Linguistics*, 4:17–30.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2020. DagoBERT: Generating derivational morphology with a pretrained language model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3848–3861, Online. 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.

Valentin Hofmann, Leonie Weissweiler, David Mortensen, Hinrich Schütze, and Janet Pierrehumbert. 2024. Derivational Morphology Reveals Analogical Generalization in Large Language Models. *arXiv preprint*. ArXiv:2411.07990.

Itay Itzhak and Omer Levy. 2022. Models In a Spelling Bee: Language Models Implicitly Learn the Character Composition of Tokens. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5061–5068, Seattle, United States. Association for Computational Linguistics.

Taku Kudo. 2018. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Paul Lerner and François Yvon. 2025. Towards the Machine Translation of Scientific Neologisms. In *Proceedings of the 31st International Conference on Computational Linguistics*. International Committee on Computational Linguistics.

Rochelle Lieber. 2010. *Introducing morphology*. Cambridge University Press, Cambridge. OCLC: 650278652.

Meta Llama Team. 2024. The Llama 3 Herd of Models.

David R. Mortensen, Valentina Izrailevitch, Yunze Xiao, Hinrich Schütze, and Leonie Weissweiler. 2024. Verbing weirds language (models): Evaluation of English zero-derivation in five LLMs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17359–17364, Torino, Italia. ELRA and ICCL.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting Linguistic Patterns in Human and LLM-Generated News Text. *Artificial Intelligence Review*, 57(10):265.

Boris New, Jessica Bourgin, Julien Barra, and Christophe Pallier. 2024. UniPseudo: A universal pseudoword generator. *Quarterly Journal of Experimental Psychology*, 77(2):278–286. Publisher: SAGE Publications.

Byung-Doh Oh and William Schuler. 2024. Leading whitespaces of language models' subword vocabulary pose a confound for calculating word probabilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3464–3472, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint*. ArXiv:2303.08774 [cs].

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*, 32.

Mohammad Taher Pilehvar. 2019. On the importance of distinguishing word meaning representations: A case study on reverse dictionary mapping. 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 2151–2156, Minneapolis, Minnesota. Association for Computational Linguistics.

Tiago Pimentel and Clara Meister. 2024. How to compute the probability of a word. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18358–18375, Miami, Florida, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21:1–67.

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.

Pavol Štekauer. 2012. *Word-formation in the World's Languages: A Typological Survey*. Cambridge University Press.

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. In *ICLR 2024*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint*. ArXiv:2307.09288 [cs].

Thinh Truong, Yulia Otmakhova, Karin Verspoor, Trevor Cohn, and Timothy Baldwin. 2024. Revisiting subword tokenization: A case study on affixal negation in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5082–5095, Mexico City, Mexico. Association for Computational Linguistics.

Julie Tytgat, Guillaume Wisniewski, and Adrien Betrancourt. 2024. Évaluation de la Similarité Textuelle : Entre Sémantique et Surface dans les Représentations Neuronales. In *35èmes Journées d'Études sur la Parole (JEP 2024) 31ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2024) 26ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2024)*, pages 85–96, Toulouse, France. ATALA & AFPC.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv:1609.08144 [cs]*. ArXiv: 1609.08144.

Armel Zebaze, Benoît Sagot, and Rachel Bawden. 2024. In-Context Example Selection via Similarity Search Improves Low-Resource Machine Translation. *arXiv preprint*. ArXiv:2408.00397.

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 and virtual meeting. Association for Computational Linguistics.

A Rules of Morphotactics

The following rules were used to create controlled datasets pairing a base (e.g., "*lightfast*") with a prefixed derivative (e.g., "*unlightfast*") and a suffixed derivative (e.g., "*lightfastly*"; see Section 3.5).

For English (i) The base should not start with "*un*" to avoid a double negation. (ii) The base should not end with:

- "*y*" because it would then have to be substituted by "*i*" (as in "*easy*" → "*easily*");
- "*le*" because it would be deleted (as in "*noble*" → "*nobly*");
- "*ll*" because the suffix would then be "*-y*" instead of "*-ly*" (as in "*full*" → "*fully*") ;
- "*ic*" to avoid allomorphy with the suffix"*-ally*" (as in "*allergic*" → "*allergically*").

For French (i) The base should not start with:

- "*i*" to avoid a double negation;
- "*b*", "*l*", "*m*", "*n*", "*p*", or "*r*" to avoid allomorphy with the "*i-*" prefix (also respectively written "*il-*", "*im-*", or "*ir-*"), as in "*irréaliste*".

(ii) The base *should* end with an "*e*" so that the "*-ment*" suffixation is morphotactic (e.g., avoid impossible words like "**absentment*") and orthographic (e.g., adverbs are often formed on the feminine adjectival form that ends with an "*e*": "*amicalement*" and not "**amicalment*").

B Complete Results

Table 4 reports the scores that are plotted in Figure 2.

C Implementation Details

LLMs are implemented in the `transformers` library (Wolf et al., 2020) itself based on `pytorch` (Paszke et al., 2019). LLMs are quantized in 8 bits for effective inference on a single V100 GPU with 32GB of RAM. We use greedy decoding.

Model	Dataset	Random ICL + BPE Seg.		Morph. ICL + BPE Seg.		Morph. (ICL + Seg.)	
		Prefix	**Suffix**	**Prefix**	**Suffix**	**Prefix**	**Suffix**
CroissantLLM-1.3B	ADJ-EN	0.196	0.581	0.836	0.842	0.988	0.988
	ADJ-FR	0.047	0.356	0.371	0.754	0.991	0.947
	PSEUDO-EN	0.265	0.651	0.811	0.866	0.987	0.980
	PSEUDO-FR	0.045	0.285	0.367	0.586	0.978	0.961
Llama-2-7B	ADJ-EN	0.715	0.812	0.949	0.879	0.990	0.999
	ADJ-FR	0.549	0.565	0.889	0.717	0.994	0.990
	PSEUDO-EN	0.792	0.802	0.943	0.900	0.997	0.997
	PSEUDO-FR	0.597	0.390	0.868	0.523	0.988	0.988
BLOOM-560M	ADJ-EN	0.105	0.477	0.561	0.771	0.998	0.996
	ADJ-FR	0.005	0.230	0.055	0.457	0.970	0.993
	PSEUDO-EN	0.077	0.556	0.535	0.772	0.996	0.992
	PSEUDO-FR	0.002	0.141	0.028	0.275	0.969	0.981
BLOOM-1.1B	ADJ-EN	0.154	0.412	0.595	0.696	0.995	0.996
	ADJ-FR	0.063	0.166	0.280	0.285	0.978	0.985
	PSEUDO-EN	0.175	0.632	0.578	0.783	0.962	0.951
	PSEUDO-FR	0.037	0.087	0.254	0.217	0.981	0.981
BLOOM-3B	ADJ-EN	0.251	0.663	0.796	0.860	0.998	0.995
	ADJ-FR	0.132	0.448	0.556	0.635	0.994	0.987
	PSEUDO-EN	0.298	0.663	0.759	0.785	0.997	0.995
	PSEUDO-FR	0.106	0.292	0.516	0.470	0.988	0.981
BLOOM-7.1B	ADJ-EN	0.488	0.724	0.893	0.879	0.999	0.998
	ADJ-FR	0.262	0.560	0.655	0.746	0.995	0.996
	PSEUDO-EN	0.625	0.799	0.873	0.868	0.999	0.998
	PSEUDO-FR	0.278	0.435	0.670	0.585	0.998	0.994

Table 4: Numbers in Figure 2