Assessing the Morphological Competence of LLMs

M2 Internship

Keywords: Large Language Models, Morphology, Natural Language Processing

Research Context and Questions

Hofmann et al. (2025) studied the modeling of competition between the nominal suffixes *-ity* (*available* → *availability*) and *-ness* (*selfish* → *selfishness*). They find that LLMs model this competition fairly well. However, they did not study the competition between prefixes (e.g., *un-* and *non-*).

This is an important research question because studying the morphological competence of LLMs allows us to measure their generalization ability (Weissweiler et al., 2023; Weller-Di Marco and Fraser, 2024; Lerner and Yvon, 2025b). Indeed, the lexicon is not a list of words that is known a priori and immutable (Corbin, 2012; Štekauer et al., 2005). However, LLMs are probabilistic models trained to maximize the likelihood of their training data. They model a probability distribution over a finite token vocabulary. While infrequent words in a corpus were typically filtered out in traditional approaches (Eisenstein, 2019), modern models (OpenAI, 2023; Llama Team, 2024; Gemma Team, 2024) all rely on BPE (*Byte Pair Encoding*) segmentation, which segments rare words into subwords by optimizing a data compression criterion (Gage, 1994; Sennrich et al., 2016; Beinborn and Pinter, 2023). Thus, models are theoretically capable of deriving or inflecting words in forms absent from their training corpus, but the reality is more complex (Hofmann et al., 2020; Weissweiler et al., 2023; Lerner and Yvon, 2025b). Morphologically competent LLMs would be useful for a wide range of NLP applications, notably for Machine Translation (Ataman et al., 2019; Marco et al., 2022; Lerner and Yvon, 2025a), and more generally for Natural Language Generation.

Our previous work is limited to fairly simple concatenative phenomena (e.g., the prefixation of *pré+entraînement*). However, several affixes can be competitive/synonymous (Corbin, 2012), for example *pré-* and *anté-*, which raises the following question: given the same definition, could we produce *antéentraînement* rather than *préentraînement*? If not, why? It may be because of: a phonological constraint

(e.g., the number of syllables (Plénat, 2009; Lindsay and Aronoff, 2013), euphony (Lignon and Plénat, 2009)), lexical consistency (e.g., analogy with *prétraitement*), or simply historical chance (e.g., influence of English's *pretraining* (Lignon and Plénat, 2009; Holeš, 2023))?

Objectives

These questions will be assessed by comparing the probability that LLMs assign to different affixes for pseudo-words (e.g. generated using UniPseudo (New et al., 2024)) to that of a cognitively plausible model, GCM (Nosofsky, 1990). If these results are not conclusive, we will conduct a survey with native speakers, to collect judgments of acceptability (comparing, e.g. "unwug" vs. "nonwug"), in the same fashion as Hofmann et al. (2025); Copot and Bonami (2024).

Other phenomena in derivational morphology raise similar questions (Corbin, 2012), notably allomorphy, where different variants of the same morpheme can be used according to morphophonological constraints (e.g., *indétruisable vs. indestructible or, conversely, traduisible vs. *traductible).

These questions will be studied by comparing BPE-based LLMs with byte-based LLMs, which are an emerging alternative to BPE (Wang et al., 2024; Zuo et al., 2024). However, they must process much longer sequences (since a word is typically segmented into many characters or bytes), which limits the use of Transformers, whose complexity is quadratic with respect to the sequence length (Vaswani et al., 2017). This will allow us to understand why BPE-based LLMs sometimes fail or succeed in deriving new lexemes. Marco and Fraser (2024) found that for inflection, the most important criterion was the consistency of the tokenization among all inflections of a given lexeme.

Internship conditions

The internship will be supervised by Paul Lerner[1], postdoc researcher, Leonie Weissweiler[2], postdoc researcher, and François Yvon[3], senior researcher. The internship may lead to a PhD thesis, provided available funding. The internship will take place at ISIR in the MLIA team[4]. ISIR is under the dual supervision of Sorbonne University, which is a world-class multidisciplinary university, and the French National Centre for Scientific Research (CNRS), which is one of the most important research institutions in the world. ISIR includes 6 research teams and 226 people. The intern will be located at *4, place Jussieu, 75005 Paris*.

[1] https://paullerner.github.io/
[2] https://leonieweissweiler.github.io/
[3] https://fyvo.github.io/
[4] https://www.isir.upmc.fr/teams/mlia/presentation-mlia/?lang=en

- Remuneration: around 600€/month along with the refund of 75% of the Navigo (public transport) card.
- Starting date: the internship is expected to start in February or March 2026.
- Duration: 5-6 months.

Requirements

We are looking for a second-year Master's student with a strong background in Natural Language Processing/Computational Linguistics. The intern is expected to be proficient in programming, especially in the Python language, and to have already worked under Linux. They should also have experience with a deep learning framework, preferably PyTorch.

Application

Please send a resume along with a cover letter (in French or English) and grade transcripts for the last two years to Paul Lerner at `lerner@isir.upmc.fr`. A list of pointers to example projects (e.g., via GitHub) or a letter of recommendation is a plus.

References

Duygu Ataman, Wilker Aziz, and Alexandra Birch. 2019. A Latent Morphology Model for Open-Vocabulary Neural Machine Translation. In *International Conference on Learning Representations*.

Lisa Beinborn and Yuval Pinter. 2023. Analyzing cognitive plausibility of subword tokenization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4478–4486, Singapore. Association for Computational Linguistics.

Maria Copot and Olivier Bonami. 2024. Baseless derivation: The behavioural reality of derivational paradigms. *Cognitive Linguistics*, 35(2):221–250.

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

Jacob Eisenstein. 2019. *Introduction to natural language processing*. The MIT Press.

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:2408.00118.

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, Leonie Weissweiler, David R. Mortensen, Hinrich Schütze, and Janet B. Pierrehumbert. 2025. Derivational morphology reveals analogical generalization in large language models. *Proceedings of the National Academy of Sciences*, 122(19):e2423232122.

Jan Holeš. 2023. Compétition entre poly-, pluri- et multi- dans les néologismes officiels français. *Kalbotyra*, (76):42–53.

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

Paul Lerner and François Yvon. 2025b. Unlike "Likely", "Unlike" is Unlikely: BPE-based Segmentation hurts Morphological Derivations in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*. International Committee on Computational Linguistics.

Stéphanie Lignon and Marc Plénat. 2009. Echangisme suffixal et contraintes phonologiques. In Bernard Fradin, Françoise Kerleroux, and Marc Plénat, editors, *Aperçu de Morphologie Du Français*, pages 65–81. Presses Universitaires de Vincennes, Paris.

Mark Lindsay and Mark Aronoff. 2013. Natural selection in self-organizing morphological systems. *Morphology in Toulouse: Selected Proceedings of Décembrettes*, 7:133–153.

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

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

Marion Weller-Di Marco, Matthias Huck, and Alexander Fraser. 2022. Modeling Target-Side Morphology in Neural Machine Translation: A Comparison of Strategies.

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.

Robert M. Nosofsky. 1990. Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical Psychology*, 34(4):393–418.

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

Marc Plénat. 2009. Les contraintes de taille. *Aperçus de morphologie du français. Saint-Denis: Presses universitaires de Vincennes*, pages 47–63.

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, Rochelle Lieber, Marcel Den Dikken, Liliane Haegeman, Joan Maling, Guglielmo Cinque, Carol Georgopoulos, Jane Grimshaw, Michael Kenstowicz, Hilda Koopman, Howard Lasnik, Alec Marantz, John J. McCarthy, and Ian Roberts, editors. 2005. *Handbook of Word-Formation*, volume 64 of *Studies in Natural Language and Linguistic Theory*. Springer Netherlands, Dordrecht.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Junxiong Wang, Tushaar Gangavarapu, Jing Nathan Yan, and Alexander M Rush. 2024. Mambabyte: Token-free selective state space model. In *First Conference on Language Modeling*.

Leonie Weissweiler, Valentin Hofmann, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle, Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, Haofei Yu, Hinrich Schuetze, Kemal Oflazer, and David Mortensen. 2023. Counting the Bugs in ChatGPT's Wugs: A Multilingual Investigation into the Morphological Capabilities of a Large Language Model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6508–6524, Singapore. Association for Computational Linguistics.

Marion Weller-Di Marco and Alexander Fraser. 2024. Analyzing the Understanding of Morphologically Complex Words in Large Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1009–1020, Torino, Italia. ELRA and ICCL.

Jingwei Zuo, Maksim Velikanov, Dhia Eddine Rhaiem, Ilyas Chahed, Younes Belkada, Guillaume Kunsch, and Hakim Hacid. 2024. Falcon Mamba: The First Competitive Attention-free 7B Language Model.