On Obtaining Word-Level Probabilities from Subword Predictions with Masked Language Models

Anonymous ACL submission

Abstract

Masked language models output probability distributions over subword tokens, making it non-trivial to obtain reliable word-level probabilities — particularly when comparing words of different tokenization lengths. We present a systematic study of word-level probability estimation from MLMs, comparing multiple methods for probability calculation in two languages, multiple model families, and five tokenization lengths. Our most important finding is that the masking strategy used in pretraining strongly determines probability quality: models trained without whole-word masking show severely distorted within-length distributions, being outperformed even by a trigram language model for multi-token words. Among the methods we evaluate, geometric mean aggregation is the most robust and practical choice.

1 Introduction

Word-level probability estimates are a desirable output of language models, with uses in word prediction, readability assessment, acceptability judgments, model probing and lexical substitution (Ettinger, 2020; de Varda and Marelli, 2022; Kokosinskii and Arefyev, 2024). However, modern language models operate at the subword level (Sennrich et al., 2016), and this creates a fundamental comparability problem: a word tokenized into $n > 1$ subword tokens represents a qualitatively different input to the model than a single-token word, and naïve addition of token log-probabilities yields scores that are biased towards single-token words and thus not comparable across words of different tokenization lengths. This limitation affects any application relying on comparing word probabilities in context, and is particularly relevant for multilingual models and tokenizers with restricted vocabularies, which segment words more aggressively (Rust et al., 2021).

In this paper we focus on Masked Language Models (MLMs) (Devlin et al., 2019), which, contrary to autoregressive models, base their predictions on the full left and right context and are therefore especially relevant for applications in lexical semantics that require access to the full context, such as lexical anomaly detection.

We investigate multiple approaches to word-level probability estimation from MLMs, asking whether they yield comparable and reliable probabilities across tokenization lengths. While the product of token probabilities is naturally length-biased, the geometric mean – a natural normalizing alternative – is not a standard or common approach, and despite some examples of it not helping or not making a difference (Jiang et al., 2020; Bunescu and Uduehi, 2022; Ueda et al., 2024), it has not been systematically studied and compared with the product for different tokenization lengths. We therefore evaluate a range of methods: the product and geometric mean, word-constrained variants that restrict predictions to tokens that can make structurally valid words, a Bayesian decomposition that applies a length prior to reweight words by their tokenization length, and variants of pseudo-log-likelihood (Salazar et al., 2020).

All methods are evaluated on a cloze task (Taylor, 1953) and a syntactic preference task for English and French, including multiple model families pretrained with different masking strategies, with further analysis on a lexical substitution dataset.

Our experiments reveal that the masking strategy used during pretraining is by far the most consequential factor: models trained with whole-word masking produce dramatically better calibrated probability distributions, while the probabilities of token-masked models show severe distortions for multi-token words of the same length which none of the tested aggregation methods can possibly correct. Ultimately, we find that using the mean of log-probabilities serves as the most effective choice. We release the code and data used for

reproducibility and to facilitate future work.[1]

2 Related Work

Scoring individual words has been motivated by applications such as surprisal-based reading time modeling (Goodkind and Bicknell, 2018) or semantic acceptability (Temerko et al., 2025). However, extending a subword-based output distribution to words is non-trivial. For autoregressive models, Pimentel and Meister (2024) and Oh and Schuler (2024) show how to carefully calculate a word's probability distinguishing tokens that mark word boundaries. Neither work aims to compare probabilities of different words or to produce a probability distribution over a large word vocabulary. Doing so with an autoregressive model would require a prohibitive number of forward passes to account for all possible token combinations.[2] We instead focus on masked language models, which allow for obtaining probability estimates for mid-word tokens without prefilling initial tokens for each candidate, and which provide a complementary view to left-to-right probabilities that can take into account the full sentential context.

Works extracting word-level probabilities from MLMs typically resort to simplifications: A common approach is to restrict prediction to single-token words, which comes at the cost of vocabulary coverage (Petroni et al., 2019; Ettinger, 2020; Haley, 2020; Arefyev et al., 2020; Gonen et al., 2020; Lacerra et al., 2021; Michaelov et al., 2022; de Varda and Marelli, 2022; Stodden et al., 2023). Others use the first token's probability as a proxy for longer words (Bunescu and Uduehi, 2022), or propose the sum of log probabilities across tokens, without accounting for tokenization length differences (Misra, 2022; Li et al., 2021; Michaelov and Bergen, 2022). Jiang et al. (2020) explored the mean of log-probabilities to normalize multi-token word probabilities, but did not find it useful in their knowledge retrieval setting and did not investigate it further. To our knowledge, this aggregation of log-probabilities has not been systematically evaluated as a word probability estimator, which is one contribution of this work. In the context of lexical substitution, Arefyev and Bykov (2021) and Kokosinskii and Arefyev (2024) propose workarounds for multi-token words in a multilingual setting, where the problem is of greater importance due to the higher segmentation rates of multilingual tokenizers compared to their monolingual counterparts. However, the resulting scores are not mathematically comparable across words with different tokenization lengths, and one of the proposed methods requires self-supervised fine-tuning, which introduces computational cost.

A related body of work addresses the fill-in-the-blank (FITB) task, where a model must predict a span of arbitrary length at a masked position. Several works tackle this with autoregressive or seq2seq models (Donahue et al., 2020; Shen et al., 2020; Zhu et al., 2019). More closely related to our case, Kalinsky et al. (2023) address multi-token completion with MLMs. They adapt a pre-trained MLM to predict completions from a phrase vocabulary by training a small additional module. We share their motivation of going beyond single-token prediction, but differ in two key respects: we limit ourselves strictly to the word level, and we compare more lightweight methods.

Finally, at the sequence level, Salazar et al. (2020) introduced pseudo-log-likelihood scores (PLL) for MLMs, obtained by iteratively masking each token in a sentence and summing log-probabilities, and showed these yield competitive acceptability judgments. Kauf and Ivanova (2023) refined this by masking multi-token words left-to-right, ensuring the model never sees a word's continuation when predicting its tokens. When PLL is used to compare sentences differing in a single word (common in acceptability and plausibility studies with minimal pairs), differences in tokenization length are a known source of bias which is not fully mitigated with length normalization (Ueda et al., 2024). We include normalized PLL variants in this study as comparison points in our closed-candidate evaluations.

3 Probability Calculation

Subword-based MLMs are trained to predict a probability distribution over a subword vocabulary $\mathcal{S}$ for positions marked with a `[MASK]` token. Mapping these token-level probabilities to word-level probabilities is not straightforward: Words of different tokenization lengths n require different numbers of `[MASK]` tokens, which constitutes different inputs to the model and renders probability distributions incomparable. Figure 1 illustrates the effects of

[1] Upon acceptance, we will share a link to our code and data.

[2] This may be more tractable with models adapted for speculative decoding or multi-token generation (Leviathan et al., 2023; Jin et al., 2026).

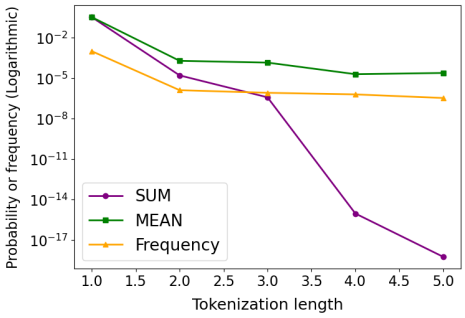

Figure 1: SUM, MEAN and frequency of correct masked words of different tokenization lengths with BERT.

using the standard sum of log-probabilities (where probability drops sharply as n increases) compared to the mean, which is more stable. However, neither measure naturally yields a valid probability distribution over words of any n. Details on how the Figure was produced are given in Appendix A.

3.1 Word probability baselines

To score a word w with tokenization $(t_1, \ldots, t_n)$, we mask all n positions simultaneously in a single forward pass.[3] We define two aggregation baselines. SUM takes the sum of log-probabilities across positions (equivalent to the product of raw probabilities). MEAN, instead, takes their mean (equivalent to the geometric mean of raw probabilities), normalizing for tokenization length.

Word-constrained variants Standard tokenizers mark word boundaries using token prefixes (e.g. BERT uses ## for continuation tokens). Similarly to Bunescu and Uduehi (2022), since in our setting we know that each masked slot contains exactly one word, we can restrict the model's predictions at each position to tokens that can validly form a word, ensuring that no probability mass is distributed over tokens that cannot structurally occur at that position. We thus define $\mathcal{S}_{\text{bow}} \subset \mathcal{S}$ as tokens that are valid at the beginning of a word, and $\mathcal{S}_{\text{mid}} \subset \mathcal{S}$ as valid continuation tokens.[4] Let $\mathbf{x}^{(n)}$ denote the input sequence where the target word is replaced by n mask tokens. We re-normalize the model's output distribution at each position m considering only the corresponding valid token subset:

$$P_{\text{bow}}(t \mid \mathbf{x}^{(n)}) = \frac{P_{\text{MLM}}(t \mid \mathbf{x}^{(n)}, m{=}1)}{\sum_{k \in \mathcal{S}_{\text{bow}}} P_{\text{MLM}}(k \mid \mathbf{x}^{(n)}, m{=}1)}, \quad t \in \mathcal{S}_{\text{bow}} \tag{1}$$

and analogously for continuation positions using $\mathcal{S}_{\text{mid}}$. Applying SUM or MEAN to these re-normalized distributions yields the word-constrained variants SUM-WC and MEAN-WC.

3.2 Vocabulary definition

To obtain a valid probability distribution, we must define the space of words that will be considered. The tokenizer's subword vocabulary $\mathcal{S}$ is not suitable as it does not entirely consist of full words. We instead pre-specify an external reference vocabulary V of valid unigrams (described in Section 4). Let $\mathcal{T}$ be a tokenizer that deterministically maps a word $w \in V$ to a sequence of subword tokens $\mathbf{t} = (t_1, t_2, \ldots, t_n)$. We partition V into disjoint sets V_n based on tokenization length: $V_n = \{w \in V : |\mathcal{T}(w)| = n\}$. For instance, V_1 contains single-token words, while V_2 contains words split into two subwords.

3.3 Bayesian decomposition with length prior

To address the length bias that occurs with SUM, we propose a Bayesian decomposition of word probability. We treat the tokenization length L as a latent variable. For any word w (tokenized into n subwords) in a context $\mathbf{c}$, we define:

$$P(w \mid \mathbf{c}) = P(w \mid \mathbf{c}, L = n) \times P(L = n \mid \mathbf{c}) \tag{2}$$

This separates the estimation problem into two distinct components: (1) the **length-conditional word probability** $P(w \mid \mathbf{c}, L = n)$, which ranks words relative to others of the same length; and (2) the **length prior** $P(L = n \mid \mathbf{c})$, which estimates how likely the masked slot is to be filled by a word of length n in this context.

3.3.1 Length-conditional word probabilities

The first component can be computed using SUM or SUM-WC, followed by normalization within each length group, yielding a valid probability distribution over every V_n separately:

$$P(w \mid \mathbf{c}, L{=}n) = \frac{P(w \mid \mathbf{c})}{\sum_{w' \in V_n} P(w' \mid \mathbf{c})} \tag{3}$$

[3] This assumes independence between the masked tokens. Progressive unmasking can instead capture inter-token dependencies, but becomes computationally prohibitive for large vocabularies. We only use it in settings with limited word candidates; see Section 3.4.

[4] The extension to tokenizers that mark the end rather than the beginning of a word is straightforward in the MLM setting, as opposed to the case of autoregressive models (Pimentel and Meister, 2024).

3.3.2 Length prior

The motivation behind the length prior is that although we can't directly trust SUM to properly predict the likelihood of an n in context, we can use other signals from the model predictions themselves to derive this information. We explore three[5] instantiations of $P(L = n \mid \mathbf{c})$:

Uniform (UNI) A context-unaware baseline assuming always equal probability for each length: $P_{UNI}(L = n \mid \mathbf{c}) = \frac{1}{|N|}$, where N is the set of considered tokenization lengths. It is therefore a simplification of $P(L = n \mid \mathbf{c})$ into $P(L = n)$. This tests whether any reweighting is beneficial at all.

Word-constrained slot probability (WCSP)**.** This prior estimates the probability that the n masked positions are filled by a single valid word, using the structural word-boundary knowledge encoded in $\mathcal{S}_{\text{bow}}$ and $\mathcal{S}_{\text{mid}}$. For each length n, we aggregate the probability mass assigned to valid tokens at each position m and take the geometric mean to normalize for length:

$$\text{WCSP}(\mathbf{x}^{(n)}) = \left(\sum_{t \in \mathcal{S}_{\text{bow}}} P(t \mid \mathbf{x}^{(n)}, m{=}1) \times \prod_{m=2}^{n} \sum_{k \in \mathcal{S}_{\text{mid}}} P(k \mid \mathbf{x}^{(n)}, m) \right)^{\frac{1}{n}} \quad (4)$$

The prior is then obtained by normalizing over all considered lengths N:

$$P_{\text{WCSP}}(L{=}n \mid \mathbf{c}) = \frac{\text{WCSP}(\mathbf{x}^{(n)})}{\sum_{j \in N} \text{WCSP}(\mathbf{x}^{(j)})} \quad (5)$$

Nucleus prior (NUC) This prior is derived from the within-n word probability distributions themselves. The intuition is that a length n is more likely when the model concentrates its probability mass on a small number of words of that length, i.e., when the distribution over V_n is peaky rather than flat. Inspired by nucleus sampling (Holtzman et al., 2020), for each n, we compute the minimum number of words k_n needed to cover $p\%$ of the probability mass over V_n (with words sorted in descending probability). A small k_n relative to $|V_n|$ indicates high confidence in length n. We convert this into a weight r_n and we normalize to obtain a valid prior:

$$r_n = \frac{|V_n|}{k_n} \qquad P_{\text{NUC}}(L{=}n \mid \mathbf{c}) = \frac{r_n}{\sum_{j \in N} r_j} \quad (6)$$

We use $p = 0.8$ in our experiments, and leave further tuning for future work.

Note that all proposed methods require exactly $|N|$ forward passes per instance, regardless of context or vocabulary size.

3.4 Progressive unmasking: PLL

For comparison, we include two more expensive scoring methods that condition on partial word observations: Pseudo-log-likelihood (PLL), as refined by Kauf and Ivanova (2023) with progressive unmasking and described in Section 2; and a word-level variant (PLL-W) that scores only the target word's tokens rather than the full sentence. For both variants, we report a raw version (sum of log-probabilities) and a version normalized by the number of tokens (PLL-NORM and PLL-W-NORM). These methods require at least one forward pass per token of the scored word, and PLL additionally scales with sentence length, so we only use them in evaluations with a limited set of candidates.

4 Experimental Setup

Vocabulary We build a reference vocabulary V for each language using the `wordfreq` library (Speer, 2022), taking the top 100k words ranked by frequency, excluding words containing digits and words tokenized into more than 5 tokens by any of our models. Figure 2 shows the resulting distribution of tokenization lengths per tokenizer.

Models For English, we use bert-large-uncased (BERT) and its whole-word-masking counterpart (BERT-WWM), which share the same 30k WordPiece vocabulary and parameter count, allowing a direct comparison. We additionally include Albert-large-v2 (Lan et al., 2020), trained with n-gram masking (Joshi et al., 2020) and SentencePiece (Kudo and Richardson, 2018), as an uncased representative of span-based masking. For French, we use FlauBERT-base-cased (Le et al., 2020), trained with token masking, and CamemBERT-v1 (Martin et al., 2020), trained with WWM. We also include mBERT as a multilingual baseline; and the results

[5] Inspired by Kalinsky et al. (2023) we also experimented with a supervised prior, training a small MLP on frozen MLM hidden states to predict a context-aware probability distribution over tokenization lengths. This was abandoned because the task was highly ambiguous, the model was highly sensitive to training regime variations, and we preferred to propose lightweight approaches.

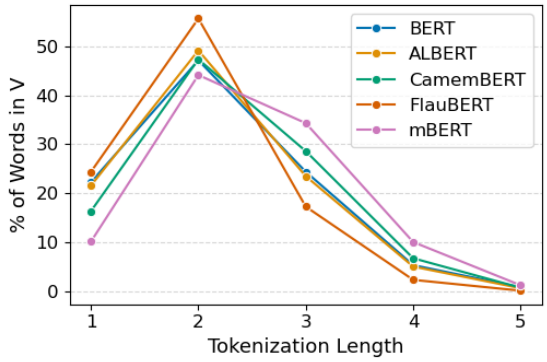

Figure 2: Percentage of words in V of each tokenization length. FlauBERT splits the least, and mBERT the most.

of another multilingual model, XLM-R (Conneau et al., 2020), are reported in Appendix D for comparison. A more detailed description of these models is provided in Appendix B.

Baseline We train a trigram model on 6.4M documents from an English Wikipedia 2023 dump with the same 100k vocabulary and Kneser-Ney smoothing, using KenLM (Heafield, 2011), which serves as a simple non-neural baseline (3-GRAM). We choose Wikipedia to approximate BERT's training data. With this model, the probability of a word corresponds to the probability of the full sentence with the target word substituted in. 3-GRAM is evaluated across n based on BERT's tokenization.

4.1 Cloze task

Our main evaluation is a cloze task where models rank all words in V for a masked slot, and we measure how highly the original word is ranked using the Mean Reciprocal Rank (MRR). For English, we use the Brown corpus (Francis and Kučera, 1967), chosen for its genre diversity. For French, we use sentences from the French portion of XNLI (Conneau et al., 2018) as provided in the FLUE benchmark (Le et al., 2020); XNLI consists of sentences from various domains that were professionally translated from English. We derive cloze instances automatically, masking one content word per sentence, avoiding stop words and filtering out sentences with less than 10 words. We do not reuse any sentence, and to ensure coverage of different tokenization lengths, we avoid repeating the same word for $n \leq 3$, and allow up to 2 occurrences of a word for $n > 3$. To allow fair comparisons between models, we balance the distribution of n across the different tokenizers where possible. This results in 16,636 English and 18,427 French instances; table 1 shows the complete dataset statistics by n and tokenizer. Figure 3 presents examples of the data used in the study.

Model	1	2	3	4	5
BERT	6,000	5,000	3,870	1,551	215
Albert	6,000	5,722	3,289	1,415	210
CamemBERT	6,013	5,000	5,000	2,000	414
FlauBERT	7,991	5,000	4,516	841	79
mBERT	5,000	5,000	5,015	2,912	500

Table 1: Statistics of the cloze datasets per tokenizer.

4.2 Syntactic preference task

This task tests whether models prefer a syntactically correct but longer word over a shorter but incorrect alternative. We augment each cloze instance of a different part of speech than the target. In English, the target's PoS comes from Brown annotations; distractor PoS mismatch is verified by inserting each candidate into the sentence and tagging with NLTK (Bird et al., 2009). In French, we use spaCy (Honnibal et al., 2020). For evaluation, we check how many distractors are ranked under the correct target for every instance, and report average accuracy.

4.3 Substitute acceptability task

If multiple words are equally acceptable in a given context, we would want models to assign them similar probabilities regardless of their tokenization length. The goal is to test how biased models are when faced with multiple acceptable words. We use the Stanford Word Substitution benchmark (Lee et al., 2021), which provides human-annotated in-context substitutes. We retain substitutes with an acceptability score ≥ 0.6 that are present in V as well as the original word. Since substitutes are given in lemma form, we inflect them using the `lemminflect` library, guided by the target word's PoS obtained with stanza (Qi et al., 2020). For each pair of acceptable substitutes of different tokenization length, we check their relative ranking as output by the model. A length-unbiased model would show a 50% preference rate; values below 50% indicate bias toward shorter words. After our filtering, we obtain 276 instances with 1,614 substitute candidates and 2,243 pairwise comparisons. This task is run in English and with BERT only.

Context: Paula was *saddened* about what was happening to little girls and vowed to kneel no more in Chapel.

Distractors: actor, incident, surroundings, distal

Context: (...) and I would learn that I had lost her long before Emil Malaquez translated her into a thing that can be kept, *admired*, and loved. (...)

Substitutes (inflected): admired, treasured (0.6), appreciated (0.6), valued (0.7), cherished (0.7), worshipped (0.6)

Figure 3: Examples of the data used for evaluation.

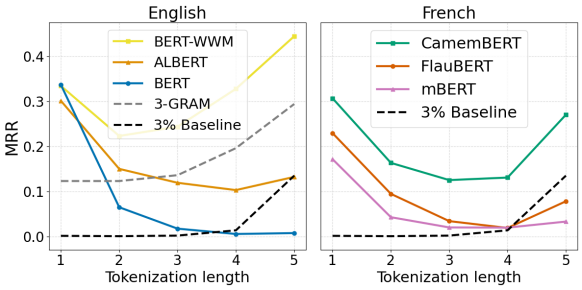

Figure 4: MRR **within** different tokenization lengths.

5 Results and Analysis

5.1 Within-length calibration

Before examining results across the full vocabulary, we assess whether models produce well-calibrated probability distributions *within* each tokenization length. Figure 4 reports MRR computed within each length group V_n,[6] together with a reference indicating the score of a model that places the target word in exactly the top 3% of V_N.[7] WWM models (BERT-WWM, CamemBERT) maintain strong MRR across different lengths, while models trained with span- and especially token-masking collapse for $n \geq 2$, even falling below the 3-GRAM model. This shows that the pretraining objective very strongly determines the quality and reliability of word probability estimation. This miscalibration is distinct from the length bias introduced by SUM, and cannot possibly be solved with length prior corrections, as they only act on between-n probability redistribution. Non-WWM models simply do not learn to reliably rank or predict words with long tokenizations.

5.2 Cloze task results

Figures 5 and 6 report MRR on the English and French cloze tasks across tokenization lengths with the full vocabulary.

[6]In this setting, SUM and MEAN are equivalent since they induce the same within-n ranking.

[7]This reference depends on $|V_n|$, which varies across tokenizers. For simplicity, we show the BERT and FlauBERT variants.

Effect of training objective While some performance degradation with higher n is expected, as longer words are rarer, we can compare how different models are affected by n. The pattern observed within-n carries over to the full cloze task: BERT-WWM maintains the strongest MRR across all n, while BERT is the most unreliable at word probability prediction even at $n = 2$, regardless of the metric used, remaining outperformed by 3-GRAM for $n \geq 2$. Albert occupies an intermediate position between the two. In French, CamemBERT (WWM) consistently outperforms FlauBERT. To further demonstrate the impact of pretraining, we include in Appendix C an experiment where continued pretraining with a WWM strategy on the BERT model yields a clear improvement on this task. We therefore recommend using WWM models whenever possible.

SUM vs MEAN and word-constrained variants SUM performs well at $n = 1$ but degrades sharply for $n > 1$, confirming the expected length bias. MEAN is substantially more robust, especially in WWM models, with only a modest cost at $n = 1$. This is expected: SUM overestimates the score of single-token words to the detriment of longer ones. Of course, single-token words are more frequent, which will cause SUM to obtain better overall accuracy in certain settings, but at the cost of misrepresenting a large part of the vocabulary. Word-constrained variants (-WC) do not yield consistent improvements over their unconstrained counterparts. Therefore, all priors are presented as calculated with the simpler, unconstrained SUM.

Length priors Our length priors generally help SUM, especially at higher n in non-WWM models, even surpassing MEAN in several cases. However, they tend to underestimate the probability of single-token words. This effect is strongest for UNI, as expected from such a context-free prior, and present to a lesser extent for NUC; while WCSP is the most robust across n. For monolingual models trained with token masking (BERT and FlauBERT), WCSP

	BERT					BERT WWM					ALBERT				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
SUM	.3366	.0016	.0002	.0000	.0000	.3314	.0674	.0398	.0462	.0529	.3010	.0214	.0097	.0052	.0049
SUM-WC	.3353	.0070	.0000	.0000	.0000	.3258	.0882	.0373	.0493	.0552	.3004	.0315	.0120	.0052	.0050
MEAN	.3321	.0100	.0028	.0002	.0001	.3024	.1171	.0934	.1271	.1803	.2916	.0615	.0353	.0199	.0134
MEAN-WC	.3190	.0238	.0008	.0002	.0001	.2826	.1427	.0896	.1308	.1742	.2800	.0789	.0391	.0193	.0129
UNI	.1707	.0241	.0061	.0012	.0010	.1601	.1039	.1348	.2120	.3238	.1746	.0673	.0598	.0558	.0753
WCSP	.3226	.0155	.0027	.0010	.0001	.2415	.0986	.1063	.2011	.3244	.2246	.0592	.0526	.0556	.0703
NUC	.2070	.0339	.0074	.0010	.0007	.1992	.1429	.1714	.1739	.1247	.2165	.0846	.0671	.0460	.0337

	FLAUBERT					CAMEMBERT					MBERT				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
SUM	.2290	.0146	.0010	.0012	.0000	.2976	.0832	.0232	.0121	.0138	.1691	.0130	.0055	.0035	.0062
SUM-WC	.2286	.0192	.0013	.0012	.0000	.2948	.0904	.0250	.0125	.0166	.1675	.0151	.0056	.0039	.0063
MEAN	.2226	.0381	.0054	.0021	.0005	.2458	.1205	.0600	.0414	.0583	.1419	.0277	.0092	.0083	.0084
MEAN-WC	.2150	.0487	.0068	.0025	.0008	.2358	.1257	.0611	.0415	.0598	.1366	.0295	.0096	.0086	.0086
UNI	.1064	.0406	.0146	.0088	.0320	.1547	.0747	.0700	.0816	.1993	.0854	.0187	.0108	.0124	.0175
WCSP	.2030	.0504	.0130	.0065	.0286	.1949	.0762	.0696	.0795	.1806	.1261	.0218	.0114	.0119	.0161
NUC	.1553	.0597	.0151	.0070	.0245	.2165	.0846	.0671	.0460	.0337	.0859	.0212	.0121	.0119	.0105

Figure 5: Cloze task results (MRR) for English and French models.

clearly outperforms MEAN on this task, but for WWM models, MEAN offers a better trade-off.

French and multilingual results French results largely reproduce the main English findings: CamemBERT (WWM) is consistently stronger than FlauBERT across all lengths and methods, and length priors help but penalize $n = 1$ more strongly than MEAN does. One notable result is that, contrary to the typical expectation that multilingual models underperform their monolingual counterparts, mBERT obtains better MRR than FlauBERT for $n \geq 3$. We hypothesize this is because multilingual training exposes the model to a higher proportion of split words, forcing the model to better learn to predict partial word fragments. mBERT is indeed the most fertile model in our vocabulary (Figure 2), while FlauBERT segments the least. This pattern is only partially reproduced by another multilingual model, XLM-R (see Appendix D), which has better results than FlauBERT at $2 \leq n \leq 3$. This could imply that, at least for models trained with token masking, quality of probability prediction is related to how frequently words of each n appear during training.

5.3 Syntactic preference results

Figure 7 shows accuracy on the 4-distractor task. We show SUM, MEAN, the best prior (WCSP), and the PLL variants. For BERT-WWM, MEAN achieves near-perfect accuracy across all n (.993–.998). SUM degrades with n; and the WCSP

	1	2	3	4	5
3-GRAM	.1200	.0400	.0390	.0320	.0420

Figure 6: MRR of the 3-GRAM model on the cloze task.

prior, despite helping BERT in the cloze task, does not consistently correct syntactic errors, suggesting that the length reweighting does not address the underlying model deficiencies that lead to the wrong syntactic preferences. For BERT, MEAN results are good, and better than 3-GRAM on this task. Regarding PLL and its variants, we observe that, analogously with SUM, raw variants degrade sharply with n, confirming the need for normalizing these metrics for tokenization length when comparing words in the same context. However, these metrics, while robust, are not as powerful as MEAN in French models.

5.4 Substitute acceptability results

Table 2 reports results on the substitute acceptability task, where a length-unbiased metric would obtain 0.5. Again, SUM shows extreme bias. This is attenuated by the MEAN, with a near-perfect balance in BERT-WWM (.455). Among PLL metrics, the normalized variants display the lowest bias; PLL-W-NORM actually reverses the bias in BERT-WWM, displaying a preference for longer words. This is likely because in progressive unmasking longer words benefit from observing partial information, which is not the case for single-token words. Overall, MEAN with a WWM model provides again

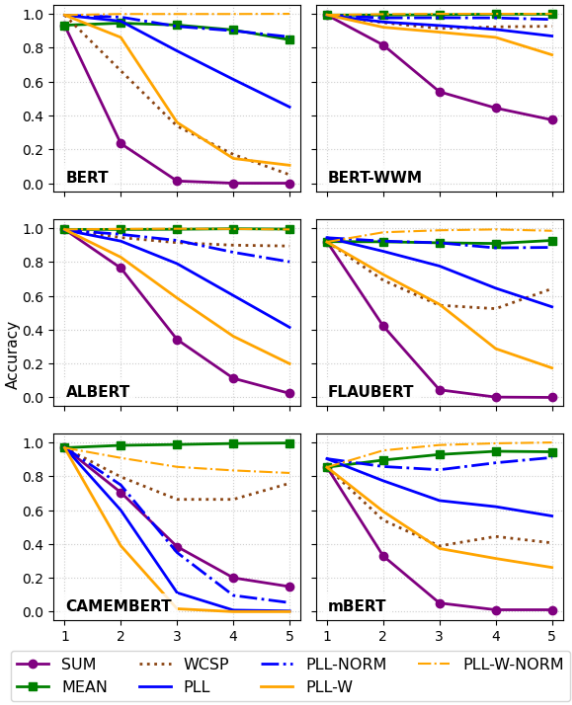

Figure 7: Accuracies on the syntactic preference task.

	BERT	BERT WWM
SUM	.005	.098
MEAN	.125	.455
WCSP	.055	.302
PLL	.125	.342
PLL-NORM	.253	.489
PLL-W	.071	.307
PLL-W-NORM	.481	.652

Table 2: Results on the substitute dataset: values lower than 0.5 indicate a bias towards shorter words.

the best balance, and would be the recommended choice for open-vocabulary substitute discovery.

6 Discussion and Conclusion

We evaluated word-level probability estimation from masked language models, comparing aggregation methods, length priors, and progressive-unmasking approaches across two languages, multiple model families, and five tokenization lengths. Pretraining masking strategy has the strongest impact on word prediction quality: models trained with whole-word masking produce dramatically better probability distributions, while other models have severely distorted distributions within longer tokenizations, being even outperformed by a trigram LM for words with $n \geq 2$ tokens. This is not completely unexpected, as the way WWM models are trained resembles our setup of masking all tokens of a word and independently predicting them. Non-WWM models may perform better with more costly partial masking approaches.

Among the evaluated methods, the mean of log-probabilities is the most robust and practical choice, as it substantially reduces length bias compared to their sum. The proposed length priors improve SUM at higher n but consistently hurt it at $n = 1$ and do not reliably outperform MEAN, as shown by their underperformance on the syntactic preference and substitute acceptability tasks. Constraining calculation to tokens that can validly form a word is not systematically beneficial either. While in settings where words with $n = 1$ dominate SUM may give better overall results thanks to its overestimation of these words, MEAN is the appropriate choice if we want to fairly consider, or compare, words with different tokenization lengths.

Even with limited candidates, as opposed to open-vocabulary settings, MEAN is more reliable across models and cheaper than length-normalized progressive unmasking (PLL-W-NORM), which may be overcorrecting thanks to partial word information. While PLL normalization is clearly beneficial in our experiments, a more extensive evaluation would be necessary for stronger generalization.

We therefore offer two main recommendations for practitioners computing word-level probabilities or derived metrics such as surprisal with MLMs: use WWM models whenever available –if no such model exists, continued pretraining with WWM is a viable alternative –; and use MEAN as the default aggregation. These recommendations are most relevant for settings involving models that split words more often, such as multilingual models, though our experiments suggest that these may be more robust to multi-token words than their monolingual counterparts, something that we plan to investigate further across more languages.

A natural direction for future work is to explore interventions on MEAN or on non-WWM models to (further) reduce length bias; and to extend this comparative study to autoregressive models, considering speculative decoding to facilitate vocabulary-wide scoring. More broadly, we hope this work brings awareness to the consequences of subword tokenization on word probability estimation, and our recommendations may prove useful for NLP tasks that rely on computing and comparing word probabilities in context, including surprisal-based modeling, lexical substitution, anomaly detection and word prediction.

Limitations

Our study focuses solely on content words, excluding proper nouns and function words; we also did not characterize words at each tokenization length with regard to their frequency, PoS, or other relevant features such as their complexity or whether their tokenization is in accordance with their morphology. Words with $n > 1$ tend to be rarer and more complex or technical, and a breakdown of results by word frequency could provide additional insight.

We did not control for different aspects of the context either, other than requiring a sentence length of at least 10 words. Context length and informativeness, and the relative position or syntactic roles of the masked words, which can have an impact on prediction quality, were not taken into account. However, we believe that the size of our datasets and our sampling method mitigate this concern to some extent.

Our evaluation is restricted to words that are tokenized into up to 5 tokens. Words with longer tokenizations are very infrequent and were excluded from our study; it is therefore unclear whether our findings generalize beyond $n = 5$. We also did not experiment with other vocabulary sizes, which may affect methods using a length prior, as the relative sizes of V_n can influence prior estimation.

We do not address multiword expressions, as they fall outside the scope of this work, but they constitute a natural direction for future studies.

Finally, we evaluated the probability distributions based solely on the rank of the correct words, without analyzing the properties of the resulting word probability distributions, such as their entropy, which could shed more light on the failures we observed.

References

Wissam Antoun, Francis Kulumba, Rian Touchent, Éric de la Clergerie, Benoît Sagot, and Djamé Seddah. 2024. CamemBERT 2.0: A Smarter French Language Model Aged to Perfection. *arXiv preprint arXiv:2411.08868*.

Nikolay Arefyev, Boris Sheludko, Alexander Podolskiy, and Alexander Panchenko. 2020. Always Keep your Target in Mind: Studying Semantics and Improving Performance of Neural Lexical Substitution. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1242–1255, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Nikolay V Arefyev and Dmitry A. Bykov. 2021. An interpretable approach to lexical semantic change detection with lexical substitution. In *Proceedings of the International Conference on Computational Linguistics and Intellectual Technologies (Dialogue)*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc., Beijing.

Razvan C. Bunescu and Oseremen O. Uduehi. 2022. Distribution-Based Measures of Surprise for Creative Language: Experiments with Humor and Metaphor. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 68–78, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Andrea de Varda and Marco Marelli. 2022. The Effects of Surprisal across Languages: Results from Native and Non-native Reading. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 138–144, Online only. Association for Computational Linguistics.

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.

Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling Language Models to Fill in the Blanks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2492–2501, Online. 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.

W. Nelson Francis and Henry. Kučera. 1967. Computational analysis of present-day american english. In *Providence, RI: Brown University Press*.

Hila Gonen, Shauli Ravfogel, Yanai Elazar, and Yoav Goldberg. 2020. It's not Greek to mBERT: Inducing Word-Level Translations from Multilingual BERT. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 45–56, Online. Association for Computational Linguistics.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

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.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-FACTR: Multilingual Factual Knowledge Retrieval from Pretrained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959, Online. Association for Computational Linguistics.

Ziqi Jin, Lei Wang, Ziwei Luo, and Aixin Sun. 2026. MARS: Enabling Autoregressive Models Multi-Token Generation. *arXiv preprint arXiv:2604.07023*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Oren Kalinsky, Guy Kushilevitz, Alexander Libov, and Yoav Goldberg. 2023. Simple and Effective Multi-Token Completion from Masked Language Models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2356–2369, Dubrovnik, Croatia. Association for Computational Linguistics.

Carina Kauf and Anna A. Ivanova. 2023. A Better Way to Do Masked Language Model Scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 925–935, Toronto, Canada. Association for Computational Linguistics.

Denis Kokosinskii and Nikolay Arefyev. 2024. Multilingual Substitution-based Word Sense Induction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11859–11872, Torino, Italia. ELRA and ICCL.

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.

Caterina Lacerra, Tommaso Pasini, Rocco Tripodi, and Roberto Navigli. 2021. ALaSca: an Automated approach for Large-Scale Lexical Substitution. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3836–3842. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Mina Lee, Chris Donahue, Robin Jia, Alexander Iyabor, and Percy Liang. 2021. Swords: A Benchmark for Lexical Substitution with Improved Data Coverage and Quality. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4362–4379, Online. Association for Computational Linguistics.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*, pages 19274–19286. PMLR.

Bai Li, Zining Zhu, Guillaume Thomas, Yang Xu, and Frank Rudzicz. 2021. How is BERT surprised? Layerwise detection of linguistic anomalies. 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 4215–4228, Online. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

James Michaelov and Ben Bergen. 2022. The more human-like the language model, the more surprisal is the best predictor of N400 amplitude. In *NeurIPS 2022 Workshop on Information-Theoretic Principles in Cognitive Systems*.

James A Michaelov, Seana Coulson, and Benjamin K Bergen. 2022. So cloze yet so far: N400 amplitude is better predicted by distributional information than human predictability judgements. *IEEE Transactions on Cognitive and Developmental Systems*, 15(3):1033–1042.

Kanishka Misra. 2022. minicons: Enabling Flexible Behavioral and Representational Analyses of Transformer Language Models. *arXiv preprint arXiv:2203.13112*.

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.

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.

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.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models. 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 3118–3135, Online. Association for Computational Linguistics.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked Language Model Scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

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.

Tianxiao Shen, Victor Quach, Regina Barzilay, and Tommi Jaakkola. 2020. Blank Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5186–5198, Online. Association for Computational Linguistics.

Robyn Speer. 2022. rspeer/wordfreq: v3.0.

Regina Stodden, Laura Kallmeyer, Lea Kawaletz, and Heidrun Dorgeloh. 2023. Using Masked Language Model Probabilities of Connectives for Stance Detection in English Discourse. In *Proceedings of the 10th Workshop on Argument Mining*, pages 11–18, Singapore. Association for Computational Linguistics.

Wilson L. Taylor. 1953. "Cloze Procedure": A New Tool for Measuring Readability. *Journalism & Mass Communication Quarterly*, 30:415 – 433.

Anna Temerko, Marcos Garcia, and Pablo Gamallo. 2025. A Continuous Approach to Metaphorically Motivated Regular Polysemy in Language Models. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 419–436, Vienna, Austria. Association for Computational Linguistics.

Naoya Ueda, Masato Mita, Teruaki Oka, and Mamoru Komachi. 2024. Token-length bias in minimal-pair paradigm datasets. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16224–16236, Torino, Italia. ELRA and ICCL.

Wanrong Zhu, Zhiting Hu, and Eric Xing. 2019. Text infilling. *arXiv preprint arXiv:1901.00158*.

A Illustration of SUM and MEAN

To produce Figure 1, we randomly collected word usages from the English portion of the XNLI dataset involving different tokenization lengths (500 for each $n \leq 4$, and 317 for $n = 5$) and coming from sentences of at least 10 words. We input the sentences into bert-large-uncased (Devlin et al., 2019) with as many mask tokens as the original word. We calculate the probability of the true occurring words in two ways: the SUM or the MEAN of their log-probabilities. We report the average probability per n with each method, as well as the frequency, obtained with the `wordfreq` library (Speer, 2022). Frequency is not provided as a ground truth, as it is completely context-unaware, but rather as a reference to take into account differences in frequency across n. As we can see, predictions by MEAN more closely reflect word frequencies.

B Full Model Details

Table 3 summarizes the parameter counts, vocabulary sizes and tokenization algorithm of each neural model. We choose CamemBERT v1 for French instead of the more recent CamemBERT v2 (Antoun et al., 2024) because while the former is trained with whole-word-masking, the v2 model masks individuals tokens (although with a higher masking rate of 30%). For consistency, since CamemBERT only comes in cased form, we choose the cased versions of the other models used for French. Note that FlauBERT is the only model that marks end-of-word tokens instead of beginning-of-word tokens. The calculation of word-constrained (-WC) variants and the WCSP prior are calculated accordingly.

For PoS-tagging in French, the spaCy model we use is fr_core_news_md. Our choice of PoS-taggers for distractor selection and substitute inflection is based on their speed-quality trade-off.

C Continued Pretraining with Whole-Word-Masking

As explained in Section 5.2, we run continued pretraining (CPT) with a whole-word-masking objective on bert-large-uncased to see if this intervention can improve word probability prediction.

To ensure that the improvement comes from the masking policy rather than exposure to new data, we choose Wikipedia as the training corpus, as BERT was partially trained on it. We use the oldest available dump in HuggingFace, which dates from November 2023 (it is the same used to train the 3-GRAM model). 15% of tokens were selected for masking, and all subword tokens belonging to the same word were masked together. We trained the model for 20K steps, covering approximately 2.6B tokens. We used AdamW with a peak learning rate of 1×10^{-5}, linear warmup over 500 steps followed by linear decay, weight decay of 0.01, and gradient clipping at norm 1.0. The effective batch size was 256 sequences.The best checkpoint was selected via early stopping based on the validation loss (evaluated every 1,000 steps on a held-out set of 60k articles).

Results are presented in Figure 8 along with the BERT and BERT-WWM results presented in the main paper, for a more direct comparison. As expected, BERT CPT-WWM is inbetween the two: there is a notorious improvement over the initial model, but some length bias remains and performance is still not on par with that of BERT WWM. This shows that, in the absence of a WWM model, this can help improve prediction.

D XLM-R Results

Figure 9 provides results obtained by XLM-R in the French cloze task, along with the other non-WWM models for comparison: FlauBERT and mBERT. While mBERT and XLM-R do not show the same behavior (XLM-R is better at $n \leq 2$), we can see that XLM-R obtains better results than the monolingual FlauBERT on $n = 2$ and 3 with the basic SUM and MEAN. We also report the within-n results for XLM-R in Figure 10. XLM-R is better calibrated than mBERT, but both multilingual models obtain overall worse results than FlauBERT in this respect.

For reference, 4 contains the full statistics of the French cloze datasets per tokenizer (similar to Table 1 but including XLM-R); and Figure 11 shows (analogously to Figure 2 in the main paper). Similarly to mBERT, XLM-R splits more words than monolingual models.

Model	Parameters	Vocabulary	Tokenizer	Case	BOW/EOW	Masking
BERT	340M	30K	WordPiece	uncased	BOW	token
BERT WWM	340M	30K	WordPiece	uncased	BOW	word
ALBERT	17M	30K SentencePiece	uncased	BOW	span	
FlauBERT	138M	69K	BPE	cased	EOW	token
CamemBERT	110M	32K	SentencePiece	cased	BOW	word
mBERT	178M	119K	WordPiece	cased	BOW	token
XLM-R	270M	250K	SentencePiece	cased	BOW	token

Table 3: Details of all models used in the study. BOW/EOW refers to whether their tokenizer marks *beginning* (BOW) or *end*-of-word (EOW) positions.

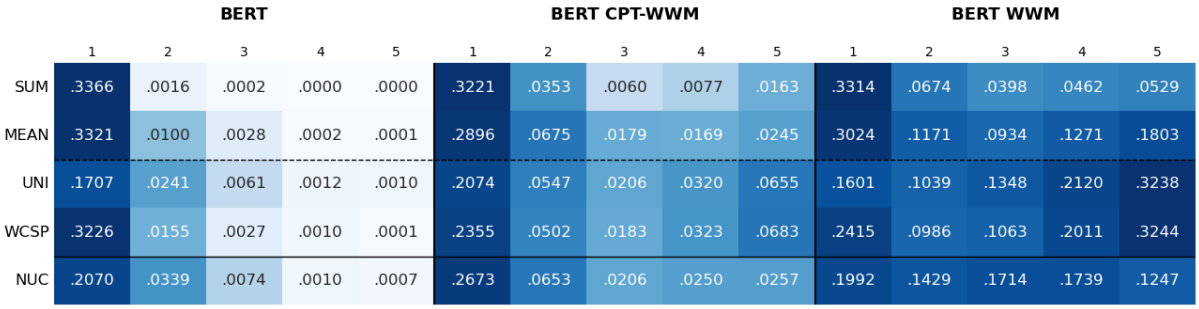

Figure 8: MRR on the cloze task of the BERT model that underwent continued pretraining with a whole-word-masking strategy, along with BERT and BERT-WWM results.

Model	1	2	3	4	5
CamemBERT	6,013	5,000	5,000	2,000	414
FlauBERT	7,991	5,000	4,516	841	79
mBERT	5,000	5,000	5,015	2,912	500
XLM-R	5,000	5,461	5,226	2,240	500

Table 4: Statistics of the French cloze dataset per tokenizer.

	FLAUBERT					MBERT					XLM-R				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
SUM	.2290	.0146	.0010	.0012	.0000	.1691	.0130	.0055	.0035	.0062	.2338	.0338	.0033	.0012	.0000
MEAN	.2226	.0381	.0054	.0021	.0005	.1419	.0277	.0092	.0083	.0084	.1994	.0585	.0087	.0053	.0002
UNI	.1064	.0406	.0146	.0088	.0320	.0854	.0187	.0108	.0124	.0175	.1383	.0389	.0108	.0122	.0058
WCSP	.2030	.0504	.0130	.0065	.0286	.1261	.0218	.0114	.0119	.0161	.1843	.0463	.0107	.0108	.0038
NUC	.1553	.0597	.0151	.0070	.0245	.0859	.0212	.0121	.0119	.0105	.1510	.0556	.0128	.0111	.0023

Figure 9: MRR on the French cloze task of FlauBERT and two multilingual models: mBERT and XLM-R.

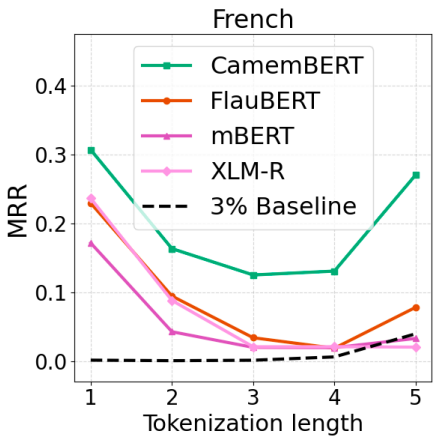

Figure 10: Within-n MRR of all models on the French cloze task.

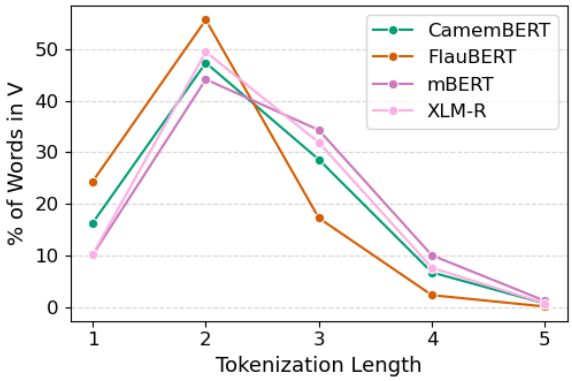

Figure 11: Percentage of words in V of each tokenization length for all models (tokenizers) used in French.