

Arabic BabyLM: Effects of Diglossia and Tokenisation

Mozafar Haider Burcu Can

School of Computing, Data, and Mathematical Sciences

University of Stirling

Stirling, Scotland

mozafar@nyala.dev

burcu.can@stir.ac.uk

Abstract

This paper explores building a baby language model for Arabic in the context of the BabyLM challenge, with a pretraining corpus limited to 10 million words. We focus on two areas: 1. morphologically-aware segmentation and tokenisation methods suitable for Arabic, and 2. the unique challenges of data curation in diglossic linguistic settings, where the language spoken at home and in everyday communication (Dialectal Arabic) differs substantially from the written language used in formal education and most media (Modern Standard Arabic). We evaluate our models using the BabyBabelLM evaluation pipeline. We achieved a significant improvement in the grammaticality scores using different tokenisation and data composition strategies. We discuss these improvements in detail, especially what they reveal about the effect of mixing Modern Standard Arabic and Dialectal Arabic in baby Arabic language models, as well as the challenges they pose for building developmentally plausible models in Arabic.

1 Introduction

The general trend in Large Language Models (LLMs) has been to scale up both model size and dataset size (Kaplan et al., 2020; Hoffmann et al., 2022). However, this leads to computationally inefficient models that are power and data hungry (Bender et al., 2021).

The contrast with how humans learn languages is stark. While humans master languages with exposure to around 100 million words (Gilkerson et al., 2017), a (very) old model like BERT (2018) (Devlin et al., 2019) is trained with 3 billion tokens, and GPT-3 (2020) was trained on 300 billion tokens (Brown et al., 2020). Modern models are trained over trillions of words, OLMo 2 for example, is an open-weight model trained on 6.6 trillion tokens (OLMo Team et al., 2025).

This substantial gap in data efficiency has motivated research on small and baby language models

with fewer parameters that can be trained on substantially less data (Lu et al., 2024; Charpentier et al., 2025).

The BabyLM challenge (Warstadt et al., 2023) addresses this gap aiming to develop data-efficient approaches that more closely mirror human language acquisition, particularly that of children. The challenge introduces a multilingual track this year in BabyLM 2026 (BabyLM Team et al., 2026; Jumelet et al., 2025) after following English-only shared tasks between 2023-2025. A developmentally-plausible Arabic dataset and evaluation pipeline were released as part of the multilingual track (BabyBabelLM).

In this paper, we explore building a baby language model for Arabic. Extending BabyLM to Arabic introduces additional challenges due to the linguistic structure of the language. First, Arabic is a morphologically rich language with complex concatenative, derivational, templatic and inflectional processes (Habash, 2010). Second, data curation introduces its own challenges because of various reasons: 1. Due to the diglossic nature of Arabic, the spoken and written languages differ substantially (Holes, 2004), creating challenges for child speech data, as children are exposed primarily to spoken Dialectal Arabic (DA) while Modern Standard Arabic (MSA) is used in written language. 2. Available corpora combine text from different dialects which differ lexically, morphologically and syntactically.

To overcome these challenges, we focus on two areas: Tokenisation and dataset curation. We explore several segmentation and tokenisation methods suitable for Arabic morphology and we study their interaction with dataset composition and their performance in different tasks categories (i.e. grammaticality, reasoning, and knowledge). The results show a strong improvement in grammatical learning over the baseline, where morphologically-aware tokenisation paired with certain dataset com-

position strategies, achieves 80% in MultiBLiMP (multilingual extension of the grammatical benchmark) (Jumelet et al., 2026).

Our main contributions are:

- A comparative and statistical analysis of various segmentation and tokenisation methods and their effects on improving Arabic language models.
- An analysis of the effects of dataset composition on downstream tasks, especially in terms of mixing Arabic dialects and MSA.
- Five new Arabic datasets along with the best performing models published on HuggingFace.¹

The implementation of the models as well as the notebooks used for analysis and dataset composition are publicly available².

2 Background and Related Work

2.1 The Arabic context

The trend towards increasingly large models, in terms of the parameter count and dataset size, has also extended beyond English to languages such as Arabic where the language modeling landscape is dominated by large headline-grabbing models (Mashaabi et al., 2024). Much of the Arabic LLM landscape is driven by large, government or semi-government-affiliated initiatives: ALLaM, developed by Saudi Arabia’s National Center for AI (Bari et al., 2024), trained over 1.2 trillion tokens (of which 45% is Arabic); Fanar, developed at Qatar Computing Research Institute, Hamad Bin Khalifa University (Fanar Team et al., 2025), trained on a corpus of 1 trillion tokens (40% Arabic, 50% English, 10% programming languages); and Falcon-Arabic, developed at the Technology Innovation Institute, Abu Dhabi (Boussaha et al., 2025), trained over 600 billion tokens (40% Arabic).

On the smaller side of models, Kuwait is one of the few models described as a Small Language Model (SLM), although it is trained with 90 billion Arabic tokens (Hennara et al., 2025). AraBERT is trained with 70 million sentences (more than 2.5 billion tokens) (Antoun et al., 2020). To the best of our current knowledge, there is no Arabic model at the scale of a baby language model.

For Arabic, a language shared by more than 20 countries with extremely different socio-economic levels, and where the distribution of wealth does not necessarily equate to the distribution of talent, this is very relevant and ties into the BabyLM’s mission of democratising research opportunities in language modeling, as it deprives many researchers from the less economically developed parts of the Arab world from the opportunity to contribute to research.

2.2 Related work

Beyond the baseline submission from BabyBabelLM (Jumelet et al., 2025), relatively little work has explored Arabic language models at this scale. However, several of the techniques investigated in this work have been studied extensively in the context of larger language models.

Tokenisation is an important topic in the Arabic NLP scene (Abdelali et al., 2016; Arnett and Bergen, 2025). This is driven by the morphological structure of Arabic that differs fundamentally from English and requires tokenisation strategies that are better suited to its rich morphology. Therefore, Arabic-aware tokenisation and segmentation is at the core of the Arabic LLMs (Boussaha et al., 2025; Fanar Team et al., 2025; Hennara et al., 2025). In this paper, we explore different techniques for Arabic tokenisation: rule-based Arabic-aware segmentation, stemming, unsupervised learning that learns the morphology automatically from a given corpus along Unigram and Byte Pair Encoding (BPE).

Dataset creation and dataset quality are crucial in SLMs in general, however, the Arabic case has its own specific challenges: the mixture and interactions between dialects and MSA (Habash, 2010; Holes, 2004), culturally-aligned corpora (Al-rashed et al., 2024) among other challenges like code-switching and orthographic inconsistencies of written Dialectal Arabic. Dataset creation has also been one of the most effective approaches in previous years’ BabyLM challenges. Prévot et al. (2024) uses speech data in French to create a linguistic corpora, Ghanizadeh and Dousti (2024) curate a dataset primarily sourced from child-directed transcripts, Velasco and Roque (2025) curate a dataset with simpler text, using shorter sentences, simpler sentence structures, and simpler vocabulary.

In the context of **BabyLM**, several studies have examined the role of morphology in language modeling for morphologically-rich languages. For ex-

¹The datasets and models are available at <https://huggingface.co/kabaro/datasets>

²The code is on: <https://github.com/kabaro/babylm-ara>

Word	Meaning
kataba	to write
maktab	desk
maktaba	library
kitab	book
kutub	books
kutayyib	booklet
maktub	written
katib	writer

Table 1: An example of the non-concatenative root-and-pattern morphology of Arabic: The root **k-t-b** (“write”) realised across multiple templatic patterns.

ample, Bölücü and Can (2025) looked at the effect of morphologically-aware tokenisation in English, Kriš and Šuppa (2025) researched building a BabyLM for Slovak which has an inflection-based morphology different from English, and Gelboim and Sulem (2025) explored learning grammatical rules in Hebrew, a Semitic language morphologically very close to Arabic. Fusco et al. (2025) also explored morphologically-aware tokenisation techniques finding that they improve comprehension generalisation at the cost of syntactic accuracy.

3 Arabic Morphology

Morphology plays a central role in Arabic NLP due to its deep interaction with both orthography and syntax (Habash, 2010, Chapter 4).

3.1 Dimensions of morphology

Arabic, similar to other semitic languages like Amharic and Hebrew, is fusional with an inflectional morphology, where consonant roots follow various templates that contribute to grammatical categories and lexical meaning. Contrast to concatenative morphology where affixes are added at the end of each word corresponding to a single grammatical feature (person, tense etc.), Arabic (like other fusional languages) has inflectional morphology and affixes can encode several grammatical categories at once.

Similarly to agglutinative languages such as Turkish and Finnish, Arabic is also considered morphologically rich. A single root can generate dozens of related word forms. This richness increases lexical sparsity and motivates morphologically-aware NLP models. Table 1 illustrates this richness with the trilateral root **k-t-b** (“write”), shown in Latin transliteration.

3.2 Does morphology matter?

Complex inflectional morphology is one of the factors contributing to differences between languages in NLP tasks (Arnett and Bergen, 2025). Morphological richness generates a high number of word forms from a single root and it substantially increases the vocabulary size in NLP applications (Erdmann and Habash, 2018). Arabic, specifically, can have out-of-vocabulary rate up to ten times higher than that of English (Habash, 2010) when using a 64K-word vocabulary.

To keep the vocabulary size optimum, morphologically-aware segmentation and tokenisation methods have been widely explored particularly in such morphologically rich languages, where they help reduce vocabulary size, improve lexical generalization, and mitigate out-of-vocabulary issues.

Tokenisation methods such as BPE and WordPiece generate subword tokens which are usually not aligned with linguistic morphemes, and this has motivated some research to better align subword tokens with linguistic morpheme boundaries (Bauwens and Delobelle, 2024). Morpheme-aware tokenisation has also been explored for Arabic language in a recent study (Asgari et al., 2025), where the results show that morpheme-aware tokenisation improves intrinsic morphological coherence and reduces language model cross-entropy, and not only in Arabic, but also in other morphologically rich languages such as Russian and Hungarian.

4 Dataset Creation for Arabic

4.1 Limitations of the BabyBabelLM corpus

BabyBabelLM provides an 8.3M dataset for Arabic³ as part of the first multi-lingual track in BabyLM 2026 (Jumelet et al., 2025). The BabyBabelLM itself discusses the limitations of this dataset such as the need to mix up different dialects into one dataset due to the lack of developmentally plausible corpora in Arabic.

We looked into possible modifications to the dataset to make it more developmentally plausible and more coherent in terms of its composition.

The current corpus is over-reliant on two sources: OpenSubtitles and the Habibi corpus (El-Haj, 2020) for Arabic songs. Together, they make almost 80% of the total number of tokens. Subtitles make up

³<https://huggingface.co/datasets/BabyLM-community/babylm-ara>

Data Source	BabyLM Arabic		Custom 8.3M blend [†]		BabyLM Spanish [‡]	
	Tokens	%	Tokens	%	Tokens	%
Habibi (Arabic songs)	3.11M	37.18%	1.39M	16.78%	103.4K	1.1%
OpenSubtitles	3.52M	42.14%	1.40M	16.82%	1.34M	13.9%
The Arabic e-book corpus	1.57M	18.84%	1.57M	18.95%	3.95M	40.7%
vikidia	4.9K	0.06%	4.9K	0.06%	1.44M	14.8%
GlottStoryBook	12.1K	0.14%	12.1K	0.15%	—	—
Ririro	81.9K	0.98%	81.9K	0.99%	—	—
MagicHub Yemeni	36.0K	0.43%	—	—	—	—
MagicHub Egyptian	18.5K	0.22%	18.5K	0.22%	—	—
other-books (e-book genres; not in original)	—	—	697.1K	8.39%	—	—
Hindawi (YA fiction; not in original)	—	—	3.13M	37.64%	—	—
child-directed-speech (Spanish-only; not in baby1m-ara)	—	—	—	—	2.87M	29.6%
Total	8.35M	100.00%	8.30M	100.00%	9.71M	100.00%

Table 2: Data source composition of baby1m-ara (Arabic dataset from BabyBabelLM) vs. our 8.3M custom blend vs. baby1m-spa (Spanish dataset from BabyBabelLM), by token count and percentage of each dataset’s own total.

[†] custom-8.3M is the 8.3M Diverse blend. [‡] baby1m-spa figures are aggregated liberally to give a general sense of comparison. Habibi equivalent in Spanish are the child-available-speech. The Arabic e-book corpus equivalent are the child-books category (with many more data sources), and child-directed-speech is a major category that does not exist in Arabic.

42% of the corpus. For context, only 13% of the Spanish corpus (a comparable Tier 2 language in the BabyBabelLM datasets) consists of subtitles.

The subtitles dataset also has other issues. To analyse it further, we augmented the dataset with information about the subtitles, using Anthropic Haiku model building a script that infers the movie titles from the subtitles, and extracts the ratings from IMDB API.

From the 559 movies in the original datasets, 347 are duplicated. For example, four versions of *Back to the Future Part III* appear in the dataset with identical translations except for the translator’s credits at the beginning. The subtitles also include a large number of movies that are not suitable for children. Table 9 in Appendix A.1 shows the breakdown of the dataset by movie age rating.

To address these issues, we rebuilt the subtitles portion of the dataset from OPUS OpenSubtitles dataset. We retrieved each movie’s title and age rating using the IMDB ID used provided in the corpus, and filtered the list to retain only the movies suitable for children. We then ranked them from the most suitable (U Rating) to PG-13, while discarding movies for adults (age rating 15, 18+). We also kept the age rating metadata as part of the augmented dataset to help further experiments in the future (e.g. for setting up age-appropriate curriculum learning).

4.2 Modified datasets

As a result of the previous analysis, we built several 8.3 million datasets with an alternative composition.

The datasets relied mostly on augmentation from the Arabic e-book corpus (Hallberg, 2024).

The original dataset already included all the children stories from the Arabic e-book corpus (which is built on freely available books from the Hindawi foundation⁴). We extended that pool with an additional subset of books from additional topics such as history, technology, arts, science, and science fiction, that we reviewed (manually) and identified as suitable for middle or high-school aged children.

We also supplemented the corpus with a popular Arabic Young Adult book series called "The Man of the Impossible"⁵. Although the series is part of the Hindawi project, it has not yet been included in the Arabic e-book corpus.

These additional texts offer several advantages. First, they are written in Modern Standard Arabic (MSA), consistent with the language used in the BabyBabelLM evaluation dataset. Second, their content is more age-appropriate than the movie-based material in the original dataset. We therefore argue that these texts more closely reflect the linguistic input to which a typical Arabic-speaking child might be exposed than the song and subtitle corpora.

While we believe that these new datasets improve upon the original dataset in terms of coherence and suitability for the evaluation tasks, the resulting corpus remains predominantly in MSA, with only a small proportion of DA content. This

⁴<https://www.hindawi.org/en>

⁵https://en.wikipedia.org/wiki/Ragol_Al_Mostaheel

distribution contrasts sharply with how Arabic is typically acquired: children are exposed predominantly, and sometimes exclusively, to DA until they reach school age (Holes, 2004). We therefore concur with Jumelet et al. (2025) on the importance of developing dialect-centric training corpora, alongside dialect-centric evaluation data, for future Arabic BabyLM research.

Table 3 presents a selection of our modified datasets and their composition. Table 2 further compares the source composition of the original dataset with one of our modified data blends and the Spanish BabyLM dataset, a comparable Tier 2 dataset from BabyBabelLM. This comparison provides context for our data composition choices while highlighting the greater diversity of sources represented in the Spanish dataset.

Name	Size	Composition
Original Data <i>BabyLM-community/babyLM-ara, unmodified</i>	8.35M	Fixed children content 20.7%, subtitles 42.1%, habibi 37.2%.
no-habibi-5M <i>Original with Habibi and other DA</i>	5.25M	Fixed children content 32.9%, subtitles 67.1%, habibi 0%. Tests effect of removing DA.
8.3M blend (ebooks-heavy) + variations <i>wiki_0-movies_x-habibi_x-ebooks_100 blend</i>	8.30M	Fixed children content 20.2%, ebooks 79.8%, (habibi movies) (0 5)%. Pure e-book augmentation (variations with 5% movies or Habibi songs).
8.3M Diverse <i>fixed-20%-movies-17%-habibi-17%-ebooks-8%-hindawi-38% blend</i>	8.30M	Fixed children content 20.2%, movies 17%, Habibi songs (Egyptian) 17%, ebooks 8%, hindawi 38% (YA books). Most evenly-spread source blend.
8.3M Impossible Man <i>fixed-20%-ebooks-42%-hindawi-38% blend</i>	8.30M	Fixed children content 20.2%, ebooks 42.2%, hindawi 37.7%. Augmented with popular Young Adult book series in MSA.

Table 3: 8M-scale modified corpora (size in words). All models contain a fixed 20% from the children books corpus.

5 Methodology

5.1 Experiments

We ran all the experiments on the same base model used in BabyBabelLM (Jumelet et al., 2025): GPT-2 light (17M parameters), trained for 10 epochs on Arabic corpora of varying composition (8.3M words similar to the original BabyBabelLM dataset). All experiments were conducted on a single NVIDIA GeForce RTX 4090 GPU. The vocabulary size was fixed at 8192. Experiments were run with three random seeds, and the reported evaluation scores are the mean of the three runs.

We deliberately kept the model architecture fixed to isolate the effects of Arabic-specific factors, fo-

cusing instead on variation in dataset composition, including content and dialectal distribution, as well as linguistic properties of Arabic, such as its morphology and syntax.

We use the multilingual evaluation suite introduced by BabyBabelLM (Jumelet et al., 2025) and report our results in the same format to facilitate direct comparison. We additionally provide a more detailed analysis of functional linguistic competence using MultiBLiMP (Jumelet et al., 2026). At the relatively small scale considered here (approximately 10M words), grammaticality judgments provide the most informative signal for Arabic, whereas performance on reasoning and knowledge-based tasks remains near chance level. We therefore focus much of our analysis on grammatical competence and its variation across different data compositions.

5.2 Tokenisation methodology

The lack of morphological alignment in tokenisation negatively impacts the performance of models for morphologically rich languages (Arnett and Bergen, 2025). Given the morphological complexity of Arabic, considerable effort has therefore been devoted to developing tokenisation and segmentation strategies that better capture its morphological structure. Fanar Team et al. (2025) created MorphBPE (Asgari et al., 2025) which incorporates Arabic morphology within BPE, and Koubaa et al. (2024) created AraNizer to address Arabic morphology in their ArabianGPT model. Allam LLM (Bari et al., 2024), Falcon (Boussaha et al., 2025) and Kuwain (Hennara et al., 2025) used tokeniser and vocabulary expansion to address the same issues of morphological misalignment in existing tokenisers. All these studies emphasise the importance of morphology-aware tokenisation in building Arabic LLMs or adapting multi-lingual LLMs to Arabic.

In this paper, we explore several tokenisation methods to compare their effectiveness across different tasks in the BabyLM challenge, particularly for the Arabic language:

- **BPE**: The most widely used tokeniser and the default baseline used by BabyBabelLM.
- **Unigram**: The Unigram model has been shown to be effective with Arabic (Kashirskiy et al., 2025) as its probabilistic approach can capture the numerous forms in Arabic morphology.
- **Farasa**: Farasa (Abdelali et al., 2016) is an SVM-based segmenter trained on Arabic-specific

lexicons and features.

- **CAMeL**: CAMeL (Obeid et al., 2020) is a library of tools for Arabic NLP. We used its specialised MSA morphological analyser for segmentation.

Farasa and CAMeL are segmenters used in pre-tokenisation. We paired them with both BPE and Unigram in our experiments.

We also explored **ISRI** (Taghva et al., 2005), an Arabic stemmer in the NLTK library, and **Morfessor** which uses unsupervised learning for morphology.

Although Morfessor (Virpioja et al., 2013) is theoretically language-independent, its segmentation approach mainly targets languages with a concatenative morphology such as Turkish. In our experiments, the results were inconsistent, and the resulting tokenisation was largely misaligned with the morphological structure of Arabic.

Our motivation for incorporating ISRI was that, although its relatively simple stemming procedure would not capture the morphological distinctions required for Arabic grammaticality judgments, its ability to identify roots might nevertheless benefit reasoning tasks. This intuition is consistent with the stemmer’s reported effectiveness in information retrieval (Taghva et al., 2005). The approach is also interesting from the perspective of developmental plausibility: we hypothesise that an early emphasis on lexical roots may loosely resemble the vocabulary-building stage of early Arabic language acquisition. Moreover, root-based representations may be more robust to dialectal variation, as many dialectal forms can be traced to roots shared with MSA (Darwish et al., 2014). Our results provide some support for the reasoning hypothesis, albeit with an important caveat: ISRI outperformed the more sophisticated segmenters on zero-shot reasoning tasks, but, as in the other experiments, its absolute performance remained only marginally above chance. We therefore omitted both ISRI and Morfessor from the main results and discussion, as neither yielded additional insights beyond these largely negative findings.

6 Results

Zero-shot grammaticality scores on MultiBLiMP are reported in Table 4. Grammaticality showed the largest improvements across our experiments, making these results particularly informative for identifying potential optimisation strategies for small-

scale Arabic language models.

We also report results on the reasoning and knowledge tasks. However, consistent with the baselines reported by Jumelet et al. (2025), performance showed no substantial improvement and remained close to chance level, even after fine-tuning.

6.1 Arabic BabyLMs can learn grammar

Across all experiments, grammaticality results improved with the dataset and tokeniser choices. The baseline score in Jumelet et al. (2025) is 75.9 (our multi-seed reproduction of the same setup is lower at 74.73) was improved significantly (up to 5%) by changing the corpus composition and the tokeniser.

Across our five datasets, the morphologically-aware tokenisers CAMeL and Farasa achieved the best scores overall, when paired with Unigram. The e-books heavy dataset was the best performing dataset. The best score achieved on these datasets was 79.23. For context, OLMo-2, a state-of-the-art 32 billion parameters model, trained over 6 trillion tokens, achieves 87% on MultiBLiMP in Arabic (Jumelet et al., 2026).

6.1.1 Tokenisation and dataset composition both affect grammatical learning

Both tokenisation and dataset composition have a significant effect on MultiBLiMP performance. The two-way ANOVA decomposition reported in Table 5 indicates that dataset composition is the stronger factor, accounting for 42.1% of the total variance, compared with 31.9% explained by tokeniser choice. Both effects are statistically significant, showing that improvements in small-scale Arabic language modelling depend on both the composition of the training data and the choice of tokenisation strategy.

Source	Sum Sq	df	<i>F</i>	<i>p</i>	% variance
Dataset	109.91	5	31.41	<0.001	42.1
Tokeniser	83.40	5	23.83	<0.001	31.9
Residual	67.89	97	—	—	26.0

Table 5: Two-way ANOVA decomposition of multiblimp_arb scores on the 6-dataset $\times$ 6-tokeniser grid from Table 4 (fit on per-seed scores, not dataset/tokeniser cell means). Both dataset and tokeniser choices are statistically significant.

Dataset	BPE	CAMeL (BPE)	CAMeL (Unigram)	Farasa (BPE)	Farasa (Unigram)	Unigram	Average by dataset
Original Data	74.73	75.33	76.43	74.57	75.33	75.14	75.25
8.3M (ebooks-heavy)	76.57	79.23	79.12	78.33	77.70	76.68	77.94
8.3M (ebooks w/ 5% habibi)	76.00	77.48	78.85	77.34	77.97	77.01	77.44
8.3M (ebooks w/ 5% movies)	75.28	77.04	78.96	77.47	77.39	75.42	76.93
8.3M (Impossible Man)	74.98	78.05	78.19	78.02	78.90	77.78	77.65
8.3M (Diverse)	74.29	75.94	77.15	74.54	76.49	75.61	75.67
Average by tokeniser	75.31	77.18	78.12	76.71	77.30	76.27	

Table 4: multiblomp scores (mean of 3 seeds) by dataset and tokeniser, with the best tokeniser per dataset in bold. Original/BPE (74.73) is our multi-seed reproduction of (Jumelet et al., 2025). The paper’s reported score is 75.9

6.1.2 Mixing Arabic varieties dilutes learning

Across all tokenisers, the highest MultiBLiMP grammaticality scores were obtained with datasets containing a high proportion of MSA. In particular, the e-books-heavy dataset achieved a mean score of 79.23% across three random seeds, with the best individual run reaching 80.3%.⁶

While non-Arabic speakers might find it reasonable, even expected, that a DA-heavy dataset would not perform well in an MSA-heavy benchmark, the notion that "it is all one Arabic" (Holes, 2004) is deeply entrenched for native Arabic speakers that we think this is a finding worth calling out explicitly.

Mixing dialects and MSA is common in larger models (Fananar Team et al., 2025; Hennara et al., 2025; Boussaha et al., 2025), but our results show that this is not beneficial for small models of BabyLM scale. Table 6 shows the effect of removing all DA content (i.e. Habibi songs, and MagicHub children conversations in Yemeni and Egyptian). Despite losing 3 million tokens, the grammaticality scores remained comparable, even improving with some tokenisers. The improvement, tellingly, happens when using vanilla BPE or Unigram rather than morphologically-aware tokenisation, further highlighting the linguistic differences between MSA and DA, where the grammar, syntax and phonology can differ substantially (Habash, 2010, ch. 1, p. 2).

6.1.3 Morphologically-aware segmenters pair better with Unigram than BPE

Pairing morphologically aware segmenters, such as Farasa and CAMeL, with Unigram tokenisation con-

	BPE	CAMeL	Farasa	Unigram
Original (8M)	74.73	75.33	74.57	75.14
No DA (5.1M)	75.12	74.78	74.21	76.00
delta	+0.39	-0.56	-0.36	+0.85

Table 6: The change in MultiBLiMP scores after removing DA content (without re-augmentation)

sistently yielded higher scores than pairing them with BPE.

Although Unigram proved effective as a standalone tokeniser without segmentation or pre-processing, we hypothesized that it would not pair well with a segmenter as segmentation would remove much of its advantage in inferring morphological boundaries. However, our experiments showed Unigram consistently outperforms BPE as a tokeniser after segmentation. Moreover, HuggingFace implementation of Unigram outperformed Google SentencePiece implementation (both as a standalone tokeniser or when paired with the segmenters).

CAMeL + Unigram performs the best for 4 out of 6 datasets, while Farasa + Unigram outperforms the others in one of them. CAMeL improves 0.88% on average, while Farasa improves 0.31% by switching from BPE to Unigram. A morphologically-aware segmentation paired with Unigram beats vanilla Unigram in all the experiments.

6.2 The tokenisers in isolation

To further investigate the effects of mixing dialects in Arabic, we analyse the tokenisers in isolation independently of their downstream language models as part of an intrinsic evaluation framework. To this end, for each tokenisation method, we used a tokeniser trained on a pure MSA dataset (the 8.3M

⁶The model corresponding to this run is publicly available at <https://huggingface.co/kabaros/ar-camel-ebooks-heavy-8.3M>.

Tokeniser	Fertility			4+ token words %			1-char tokens %			Vocab util.		
	MSA	DA	Δ	MSA	DA	Δ	MSA	DA	Δ	MSA	DA	Δ
BPE	2.39	1.99	-0.40	16.56	5.48	-11.08	41.75	19.55	-22.21	0.97	0.93	-0.04
CAMeL	2.20	2.32	+0.12	14.93	15.65	+0.72	25.97	17.10	-8.87	0.94	0.86	-0.08
CAMeL+Unigram	2.54	2.71	+0.17	22.52	25.98	+3.46	44.56	38.26	-6.30	0.96	0.89	-0.07
Farasa	1.93	2.14	+0.21	10.24	9.54	-0.69	8.19	8.86	+0.66	0.88	0.86	-0.02
Farasa+Unigram	2.63	2.80	+0.17	23.11	29.62	+6.51	51.45	49.85	-1.60	0.96	0.90	-0.06
Unigram	3.10	2.42	-0.67	23.74	16.80	-6.93	61.31	46.09	-15.22	0.97	0.89	-0.08

Table 7: Tokenisation metrics on MSA vs. DA content, using a tokeniser trained on the 8.3M Impossible Man MSA dataset. Δ is DA – MSA.

Impossible Man dataset) to tokenise the original BabyLM corpus (with its mix of MSA and DA). We then calculated tokenisation metrics following the evaluation framework that [Gazit et al. \(2025\)](#) applied to Hebrew, a language that shares Arabic templatic morphology. The metrics are calculated on the MSA part of the dataset, and the DA part separately. The metrics reported are: 1. fertility (the average number of tokens per word), 2. the percentage of words tokenised into four or more tokens or single-character tokens and 3. vocabulary utilisation indicating how much of the tokeniser’s vocabulary is being used.

The evaluation results are presented in Table 7. However, these scores alone are not reliable indicators of downstream performance in Arabic. For example, although lower fertility is generally considered desirable, morphologically rich languages may benefit from further segmenting words into meaningful subword units, resulting in higher fertility while better preserving morphological information ([Asgari et al., 2025](#)).

What the numbers show are clear patterns within tokenisation categories lines. These patterns, when combined with downstream performance, can give an indication of the intrinsic tokenisation metrics that lead to better results in the case of Arabic.

For fertility, morphologically aware tokenisers show an increase in scores from the MSA to the DA portions of the dataset, whereas vanilla BPE and Unigram show a decrease.

For 4+ token words and 1-char tokens, the percentages decrease significantly between MSA and DA for vanilla BPE and Unigram, while the morphologically-aware segmenters remain relatively stable, possibly indicating better resilience to corpora with mixed dialects that explains their better performance on mixed datasets corpora.

Unigram, whether used standalone or paired

with a segmenter, generates a higher percentage of words split into four or more tokens. While a lower percentage generally indicates better compression, which has been associated with improved downstream performance ([Gazit et al., 2025](#)), models using Unigram performed better on Arabic. One possible explanation is that its finer-grained tokenisation may better capture meaningful subword units in longer or morphologically complex words. This is in line with the findings of [Schmidt et al. \(2024\)](#), who show that better compression alone does not necessarily translate into improved downstream performance.

Vocabulary utilisation is high across tokenisers, with the exception of Farasa, which improves significantly when paired with Unigram. This provides a possible explanation for the findings reported in Section 6.1.3 where we reported that pairing with Unigram always outperforms pairing with BPE.

These scores provide additional context for the findings in Section 6.1.3, helping to explain why pairing morphologically aware tokenisation with *Unigram* yields the best overall results.

6.3 Reasoning and Knowledge

Consistent with the findings of [Jumelet et al. \(2025\)](#), the fine-tuning results are generally modest and exhibit no clear pattern. Unlike grammaticality, performance on the knowledge and reasoning tasks does not appear to respond systematically to changes in dataset composition or augmentation. This is consistent with the findings reported by [Olsson et al. \(2022\)](#) that such learning only starts in a much higher scale rather than in these tiny models.

For completeness, we provide fine-tuning results along with zero-shot results in Table 10 in Appendix A.2.

7 Conclusion and Future Work

In this work, we explore building a baby language model in Arabic. We focus on the interactions between tokenisation and dataset composition in the Arabic context, especially the interactions between Dialectal Arabic (DA) and Modern Standard Arabic (MSA).

The results show that models of this size can improve their grammatical learning despite being trained on a limited number of tokens. This finding contrasts with similar experiments on morphologically aware tokenisation in English (Fusco et al., 2025), where improvements were observed in comprehension but not in grammaticality.

We also explore the effects of mixing DA and MSA in the training corpus. We find that morphologically-aware tokenisers, paired with Unigram, perform the best. Our results confirm previous findings (and intuition) about the importance of tokenisation in morphologically-rich languages. However, consistent with previous work (Fusco et al., 2025), our results show that there is no single best approach. Rather, the interaction between dataset creation, data augmentation, and tokenisation plays a crucial role in determining the performance of language models on a small scale.

Future work could explore the trajectory of learning, both in terms of different training regimes or going beyond the 10M tokens. Other tokenisation strategies should be explored such as dedicated normalisation pipelines for Arabic orthographic peculiarities (especially for DA), as well as other techniques like splintering (Gazit et al., 2025) that are promising for Arabic. For the interaction between DA and MSA, ideas from curriculum learning where MSA can be introduced later than DA, or using age rating for an age-appropriate curriculum learning could be interesting to explore.

Cross-lingual transfer is also another potential research direction between Arabic, Amharic and Hebrew, which are typologically similar Semitic languages. However, its success likewise depends on the availability of suitable datasets in these languages.

8 Limitations

The main limitation in this study is the limited availability of child-appropriate datasets in Arabic, especially in DA, as well as the absence of suitable evaluation data for DA. The datasets we created are still far from being developmentally

plausible, even compared to other Tier 2 languages in BabyBabelLM.

For this initial exploration, we focused on providing a general overview of two topics we thought are the most important in the Arabic context (tokenisation and dataset creation). Either of these topics, in Arabic, deserves its own research and a deeper investigation.

Some of these limitations are inherited from Jumelet et al. (2025). However, like Jumelet et al., we view this work as a starting point and hope that it will open the door to further research on Arabic and provide a foundation for future work to build upon.

References

- Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. Farasa: A fast and furious segmenter for arabic. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations (NAACL-HLT)*, pages 11–16, San Diego, California. Association for Computational Linguistics.
- Sultan Alrashed, Dmitrii Khizbullin, and David R. Pugh. 2024. FineWeb-Edu-Ar: Machine-translated corpus to support Arabic small language models. *arXiv preprint arXiv:2411.06402*.
- Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resources Association.
- Catherine Arnett and Benjamin K. Bergen. 2025. Why do language models perform worse for morphologically complex languages? In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, pages 6607–6623, Abu Dhabi.
- Ehsaneddin Asgari, Yassine El Kheir, and Mohammad Ali Sadraei Javaheri. 2025. MorphBPE: A morpho-aware tokenizer bridging linguistic complexity for efficient LLM training across morphologies. *arXiv preprint arXiv:2502.00894*.
- BabyLM Team, Leshem Choshen, Ryan Cotterell, Jaap Jumelet, Tal Linzen, and Aaron Mueller. 2026. BabyLM turns 4 and goes multilingual: Call for papers for the 2026 BabyLM workshop. *arXiv preprint arXiv:2602.20092*.
- M. Saiful Bari et al. 2024. ALLaM: Large language models for Arabic and English. *arXiv preprint arXiv:2407.15390*.

- Thomas Bauwens and Pieter Delobelle. 2024. [BPE-knockout: Pruning pre-existing BPE tokenisers with backwards-compatible morphological semi-supervision](#). 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 5810–5832, Mexico City, Mexico. Association for Computational Linguistics.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. [On the dangers of stochastic parrots: Can language models be too big?](#) In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FACCT '21)*, pages 610–623, Virtual Event, Canada. ACM.
- Necva Bölücü and Burcu Can. 2025. [A morpheme-aware child-inspired language model](#). In *Proceedings of the First BabyLM Workshop*, pages 279–287, Suzhou, China. Association for Computational Linguistics.
- Basma El Amel Boussaha, Mohammed Alyafeai, Ahmed Alzubaidi, Leen AlQadi, Shaikha Alsuwaidi, and Hakim Hacid. 2025. Adapting falcon3-7b language model for arabic: Methods, challenges, and outcomes. In *Proceedings of The Third Arabic Natural Language Processing Conference*, pages 1–15. Association for Computational Linguistics.
- Tom B. 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, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.
- Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Y. Hu, Jing Liu, Jaap Jumelet, Tal Linzen, Aaron Mueller, Candace Ross, Raj Sanjay Shah, Alex Warstadt, Ethan Gotlieb Wilcox, and Adina Williams. 2025. [Findings of the third babylm challenge: Accelerating language modeling research with cognitively plausible data](#). In *Proceedings of the First BabyLM Workshop*, pages 399–417, Suzhou, China. Association for Computational Linguistics.
- Kareem Darwish, Hassan Sajjad, and Hamdy Mubarak. 2014. Verifiably effective Arabic dialect identification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1465–1468, Doha, Qatar. 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.
- Mahmoud El-Haj. 2020. Habibi – a multi dialect multi national arabic song lyrics corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 1318–1326.
- Alexander Erdmann and Nizar Habash. 2018. [Complementary strategies for low resourced morphological modeling](#). In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 54–65, Brussels, Belgium. Association for Computational Linguistics.
- Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsaneddin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, and 23 others. 2025. Fanar: An Arabic-centric multimodal generative AI platform. *arXiv preprint arXiv:2501.13944*.
- Achille Fusco, Maria Letizia Piccini Bianchessi, Tommaso Sgrizzi, Asya Zanollo, and Cristiano Chesi. 2025. Linguistic units as tokens: Intrinsic and extrinsic evaluation with BabyLM. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.
- Bar Gazit, Shaltiel Shmidman, Avi Shmidman, and Yuval Pinter. 2025. [Splintering nonconcatenative languages for better tokenization](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22405–22417, Vienna, Austria. Association for Computational Linguistics.
- Anita Gelboim and Elior Sulem. 2025. TafBERTa: Learning grammatical rules from small-scale language acquisition data in hebrew. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.
- Mohammad Amin Ghanizadeh and Mohammad Javad Dousti. 2024. [Towards data-efficient language models: A child-inspired approach to language learning](#). In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 22–27, Miami, FL, USA. Association for Computational Linguistics.
- Jill Gilkerson, Jeffrey A. Richards, Steven F. Warren, Judith K. Montgomery, Charles R. Greenwood, D. Kimbrough Oller, John HL Hansen, and Terrance D. Paul. 2017. Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology*, 26(2):248–265.

- Nizar Y. Habash. 2010. *Introduction to Arabic Natural Language Processing*. Number 10 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool.
- Andreas Hallberg. 2024. *The Arabic E-Book Corpus (version 1) [data set]*. <https://researchdata.se/en/catalogue/dataset/2024-145/1>. 81M-word corpus of 1,745 books published by the Hindawi Foundation.
- Khalil Hennara, Sara Chrouf, Mohamed Motasim Hamed, Zeina Aldallal, Mohammad Omar Hadid, and Safwan AlModhayan. 2025. Kuwain 1.5B: An arabic SLM via language injection. *arXiv preprint arXiv:2504.15120*.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, and 3 others. 2022. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*. ArXiv:2203.15556.
- Clive Holes. 2004. *Modern Arabic: Structures, Functions, and Varieties*, revised edition. Georgetown University Press.
- Jaap Jumelet, Abdellah Fourtassi, Akari Haga, Bastian Bunzeck, Bhargav Shandilya, Diana Galvan-Sosa, Faiz Ghifari Haznitrana, Francesca Padovani, Francois Meyer, Hai Hu, Julen Etxaniz, Laurent Prévot, Linyang He, María Grandury, Mila Marcheva, Negar Foroutan, Nikitas Theodoropoulos, Pouya Sadeghi, Siyuan Song, and 7 others. 2025. BabyBabelLM: A multilingual benchmark of developmentally plausible training data. *arXiv preprint arXiv:2510.10159*.
- Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2026. *MultiBLiMP 1.0: A massively multilingual benchmark of linguistic minimal pairs*. *Transactions of the Association for Computational Linguistics*, 14:193–216.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
- Mark Kashirskiy, Artiom Lipinski, and Ilya Makarov. 2025. AraToken: Optimizing Arabic tokenization with normalization pipeline and language extension for Qwen3. *arXiv preprint arXiv:2512.18399*.
- Anis Koubaa, Adel Ammar, Lahouari Ghouti, Omar Najar, and Serry Sibae. 2024. ArabianGPT: Native Arabic GPT-based large language model. *arXiv preprint arXiv:2402.15313*.
- Ľuboš Kriš and Marek Šuppa. 2025. SlovakBabyLM: Replication of the babylm and sample-efficient pre-training for a low-resource language. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.
- Zhenyan Lu, Xiang Li, Dongqi Cai, Rongjie Yi, Fangming Liu, Xiwen Zhang, Nicholas D. Lane, and Mengwei Xu. 2024. *Small language models: Survey, measurements, and insights*. *arXiv preprint arXiv:2409.15790*.
- Malak O. Mashaabi, Shahad Z. Al-Khalifa, and Hend S. Al-Khalifa. 2024. A survey of large language models for Arabic language and its dialects. *arXiv preprint arXiv:2410.20238*.
- Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhli Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMEL tools: An open source python toolkit for arabic natural language processing. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 7022–7032.
- OLMo Team, Pete Walsh, Luca Soldaini, Dirk Groeneveld, et al. 2025. 2 OLMo 2 furious. *arXiv preprint arXiv:2501.00656*.
- Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.
- Laurent Prévot, Sheng-Fu Wang, Jou-An Chi, and Shu-Kai Hsieh. 2024. *Extending the BabyLM initiative : Promoting diversity in datasets and metrics through high-quality linguistic corpora*. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 147–158, Miami, FL, USA. Association for Computational Linguistics.
- Craig W. Schmidt, Varshini Reddy, Haoran Zhang, Alec Alameddine, Omri Uzan, Yuval Pinter, and Chris Tanner. 2024. *Tokenization is more than compression*. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 678–702, Miami, Florida, USA. Association for Computational Linguistics. Accessed 9 September 2026.
- Kazem Taghva, Rania Elkhoury, and Jeffrey Coombs. 2005. Arabic stemming without a root dictionary. In *International Conference on Information Technology: Coding and Computing (ITCC'05)*, pages 152–157, Las Vegas, NV, USA.
- Dan John Velasco and Matthew Theodore Roque. 2025. *Rethinking the role of text complexity in language model pretraining*. In *Proceedings of the First BabyLM Workshop*, pages 1–28, Suzhou, China. Association for Computational Linguistics.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, and Mikko Kurimo. 2013. Morfessor 2.0: Python implementation and extensions for Morfessor Baseline. Technical Report 25/2013, Aalto University.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjape, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL 2023), Volume 2: The BabyLM Challenge*, pages 1–34. Association for Computational Linguistics.

A Appendix

A.1 Notes about the original BabyLM corpus

The original Arabic BabyBabelLM dataset relies heavily on OpenSubtitles, many movies are duplicated, shown in Table 8 and many are unsuitable for children. Table 9 shows the movie age-rating of the original corpus.

Duplicates	Sample movies
4	<i>Frankenstein, Spider-Man, The Matrix</i> and 4 more
3	<i>Shrek, Evil Dead</i> and 14 more
2	<i>The Polar Express, Rambo, Back to the Future</i> and 32 more

Table 8: Sample of duplicated movies in the original dataset (58 movies out of 211 appear more than once)

Rating	Classification	Movies	%
PG / PG-13	Child-friendly	97	45.8
G	Child-friendly	13	6.1
U	Child-friendly	9	4.2
12	Child-friendly	7	3.3
R	Adult	32	15.1
15	Adult	8	3.8
18	Adult	2	0.9
NC-17	Adult	1	0.5
(null)	Unknown	43	20.3
Total		212	100.0

Table 9: Rating composition of the OpenSubtitles movies in the original BabyLM dataset.

A.2 Complete results

Table 10 presents the complete set of results for all tasks and setups.

	Tokeniser	Zero-shot					Knowledge (fine-tuned)			Reasoning (fine-tuned)					
		<i>multitlimp</i>	<i>xcomps</i>	<i>hellaswag</i>	<i>wino grande</i>	<i>xstorycloze</i>	<i>bnlana</i>	<i>include</i>	<i>mmu</i>	<i>arc</i>	<i>belebele</i>	<i>mmi</i>	<i>truthfulqa</i>	<i>xnli</i>	<i>sib200</i>
	Random	50.0	50.0	25.0	50.0	50.0	10.0	25.0	25.0	25.0	25.0	33.0	25.0	33.0	14.3
	Paper	75.88	52.92	25.76	47.40	46.75	15.3	36.8	28.5	27.8	25.6	44.3	28.0	37.1	45.8
Original Data	BPE	74.73	51.93	25.58	50.12	47.94	13.85	36.75	28.70	27.59	27.78	46.09	31.36	41.95	47.76
	CAMEL	75.33	52.12	25.40	49.11	47.37	13.19	41.03	27.55	29.46	25.56	44.74	26.27	40.90	40.30
	CAMEL+Uni	76.43	52.08	25.78	49.27	47.34	14.18	41.03	29.40	28.63	25.56	47.10	27.12	41.40	56.22
	Farasa	74.57	51.82	25.86	50.12	48.07	13.19	35.04	27.15	26.97	28.89	45.47	29.66	42.05	52.24
	Farasa+Uni	75.33	52.89	25.77	49.60	48.40	14.34	35.90	29.15	26.56	29.44	47.72	31.36	45.20	54.23
	Unigram	75.14	52.06	25.74	49.68	47.19	15.42	40.17	27.35	26.14	33.89	48.17	26.27	43.65	55.22
8.3M (ebooks)	BPE	76.57	52.57	25.20	51.14	48.56	14.43	38.46	27.05	26.76	26.67	46.54	27.12	43.20	55.22
	CAMEL	79.23	52.32	25.31	50.35	47.97	14.51	35.90	27.25	28.22	26.11	45.81	30.51	42.40	52.24
	CAMEL+Uni	79.12	51.74	25.48	50.81	48.81	14.92	34.19	27.55	29.46	26.11	49.35	27.97	44.45	57.21
	Farasa	78.33	52.41	25.57	51.22	49.33	15.75	43.59	27.05	26.14	22.78	45.64	26.27	44.50	55.22
	Farasa+Uni	77.70	53.29	25.53	50.32	49.20	14.51	36.75	27.50	27.80	26.11	48.40	27.97	44.25	61.19
	Unigram	76.68	52.54	25.43	50.78	48.30	15.17	32.48	27.65	24.90	26.11	46.76	26.27	43.65	55.22
8.3M (ebooks w/ 5% habibi)	BPE	76.00	52.51	25.25	51.22	48.99	15.75	36.75	28.45	25.93	27.78	46.37	27.97	43.55	52.74
	CAMEL	77.48	52.34	25.13	49.93	48.04	15.50	37.61	27.65	26.97	23.33	46.43	29.66	41.95	49.75
	CAMEL+Uni	78.85	52.18	25.38	51.00	48.99	14.84	31.62	27.00	30.08	22.22	46.15	27.97	42.40	58.71
	Farasa	77.34	52.43	25.52	51.19	49.25	15.17	36.75	27.50	27.18	25.00	46.20	29.66	41.55	48.76
	Farasa+Uni	77.97	52.78	25.53	51.47	49.38	14.51	36.75	27.70	26.76	21.67	47.72	27.12	43.20	57.71
	Unigram	77.01	52.43	25.32	50.29	48.12	15.42	40.17	27.75	26.76	27.22	46.37	25.42	43.05	57.21
8.3M (ebooks w/ 5% movies)	BPE	75.28	52.59	25.27	51.47	48.17	15.50	41.03	27.00	27.80	28.33	46.60	30.51	41.50	55.72
	CAMEL	77.04	52.26	25.26	50.48	47.68	15.33	36.75	26.55	25.31	30.56	45.64	30.51	42.50	51.74
	CAMEL+Uni	78.96	52.47	25.38	51.00	48.14	15.33	32.48	26.80	25.52	25.00	47.21	35.59	44.10	51.74
	Farasa	77.47	52.09	25.49	51.77	50.13	14.43	38.46	26.70	25.93	27.22	46.88	27.97	42.50	56.72
	Farasa+Uni	77.39	53.06	25.54	51.14	49.10	14.26	36.75	28.40	28.01	26.11	48.00	31.36	43.50	56.22
	Unigram	75.42	52.21	25.52	50.21	48.12	14.34	35.90	29.40	25.93	25.56	47.27	24.58	43.05	58.71
8.3M (Impossible Man)	BPE	74.98	52.60	25.66	50.45	48.38	16.49	33.33	28.60	27.39	26.67	45.13	26.27	43.75	53.73
	CAMEL	78.05	51.73	25.60	50.84	47.39	16.82	34.19	28.85	24.27	26.67	47.55	27.12	42.70	56.72
	CAMEL+Uni	78.19	52.58	25.76	49.22	49.10	15.09	34.19	27.70	28.63	27.78	47.38	28.81	43.40	58.71
	Farasa	78.02	52.46	25.78	50.54	49.46	15.58	45.30	27.80	25.93	30.56	47.55	27.12	43.05	59.20
	Farasa+Uni	78.90	52.22	25.64	49.57	48.99	14.51	35.90	28.30	26.14	23.89	47.55	27.12	43.45	61.69
	Unigram	77.78	52.04	25.96	49.82	48.27	14.84	40.17	27.60	27.59	24.44	48.96	27.97	44.05	53.23
8.3M (Diverse)	BPE	74.29	51.97	25.73	50.51	48.53	15.25	38.46	27.95	26.14	24.44	46.43	27.97	42.45	48.26
	CAMEL	75.94	52.31	25.68	50.64	47.58	14.51	37.61	28.80	26.97	31.11	46.93	25.42	44.30	44.78
	CAMEL+Uni	77.15	52.50	25.63	49.57	48.22	15.09	35.90	27.80	26.97	29.44	47.55	28.81	41.70	60.20
	Farasa	74.54	52.77	25.73	51.34	48.45	13.77	38.46	26.75	25.73	21.11	46.31	27.12	42.95	50.25
	Farasa+Uni	76.49	53.45	25.87	50.23	48.20	15.66	34.19	28.90	28.01	23.89	48.96	28.81	44.30	63.68
	Unigram	75.61	52.01	25.88	50.10	47.86	14.43	41.88	27.80	27.59	26.11	48.11	27.97	43.25	54.23

Table 10: Full results for zero-shot and fine-tuned tasks across multiple datasets and all tokenisers (zero-shot scores are the mean of three seeds).