Tracing the Roots of Facts in Multilingual Language Models: Independent, Shared, and Transferred Knowledge

Xin Zhao[1] Naoki Yoshinaga[2] Daisuke Oba[2*]
[1]The University of Tokyo [2]Institute of Industrial Science, The University of Tokyo
xzhao@tkl.iis.u-tokyo.ac.jp {ynaga,oba}@iis.u-tokyo.ac.jp

Abstract

Acquiring factual knowledge for language models (LMs) in low-resource languages poses a serious challenge, thus resorting to cross-lingual transfer in multilingual LMs (ML-LMs). In this study, we ask how ML-LMs acquire and represent factual knowledge by conducting multilingual factual knowledge probing and a neuron-level investigation of ML-LMs. Additionally, we trace the roots of facts back to their source (Wikipedia) to understand how ML-LMs acquire specific facts. We identified three patterns in how ML-LMs acquire and represent facts: language-independent, cross-lingual shared, and transferred.

1 Introduction

To mitigate data sparsity of low-resource languages, multilingual language models (ML-LMs), such as mBERT [1] and Aya [2], are developed to facilitate knowledge transfer across languages. While cross-lingual transfer in ML-LMs has been observed in various tasks [3, 4, 5, 6, 7] due to the use of shared tokens [8, 9] and parallel corpora [10, 11, 12], previous studies have primarily concentrated on linguistic tasks like dependency parsing, and the transfer of factual knowledge remains unexplored. Previous studies using multilingual cloze-style queries for probing show that ML-LMs can recall facts across languages [13, 14, 15, 16, 17], demonstrating their ability in multilingual factual understanding. However, the mechanisms of fact representation in ML-LMs remain unclear.

We investigate whether ML-LMs exhibit cross-lingual transfer for factual knowledge with the following questions:

RQ1: Why and how does the factual probing performance of ML-LMs vary across languages? (§ 3)

RQ2: Do ML-LMs represent the same fact in different languages using a shared representation? (§ 4)

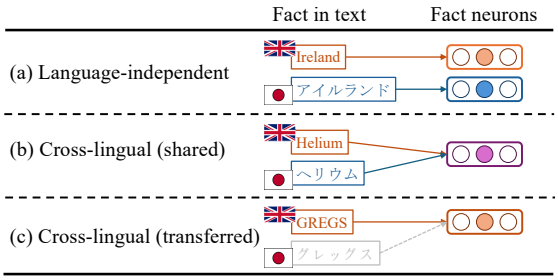

Figure 1 Three types of multilingual fact representation.

RQ3: How are cross-lingual representations of facts formed in ML-LMs during pre-training? (§ 5)

We conduct factual knowledge probing on two ML-LMs, mBERT and XLM-R, using mLAMA [14]. We reconfirm the difficulty ML-LMs face in learning facts in low-resource languages [14] (§ 2), identify factors influencing the learning of multilingual facts. We also observe that languages in geographical proximity exhibit greater overlap in shared facts, suggesting the possibility of cross-lingual knowledge transfer. Additionally, we perform a neuron-level analysis of facts to explore the role of cross-lingual transfer in fact probing. By comparing active neurons across languages, we observe that identical facts in various languages are not acquired in identical ways. Some languages share similar neuron activity for specific facts, while others exhibit distinct patterns. We categorize the former as **cross-lingual fact representations** (Figure 1(b,c)), and the latter as **language-independent** (Figure 1(a)).

To further understand cross-lingual representations, we devise a method for tracing the origins of facts by checking their presence in pretraining corpora (Wikipedia for mBERT). We assume that facts predicted correctly, even though absent in the training data, are captured through cross-lingual transfer, termed **cross-lingual transferred** (Figure 1(c)) to distinguish it from **cross-lingual shared** (Figure 1(b)). Our results reveal that only a limited number of facts can be acquired through cross-lingual transfer.

* Currently, he works for ELYZA, Inc.

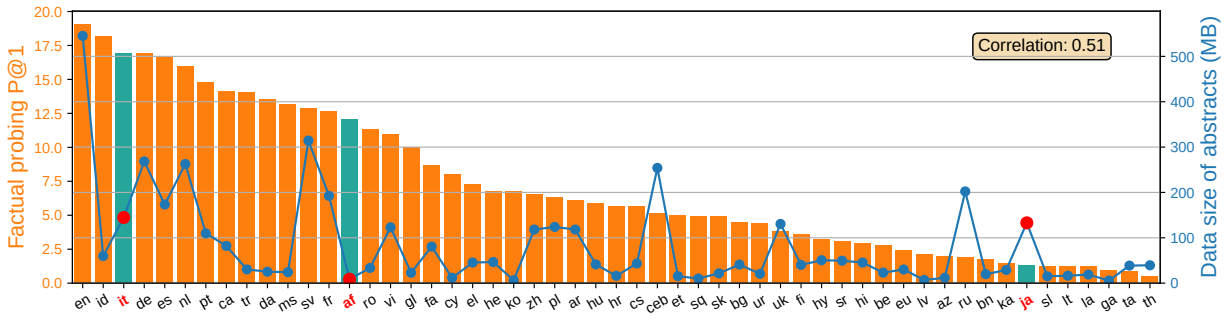

Figure 2 Wikipedia data size of abstracts vs. Factual probing P@1 on mLAMA in mBERT in 53 languages.

2 Multilingual Factual Probing

We do multilingual factual probing on ML-LMs to explore differences in factual understanding across languages.

Datasets: We use the mLAMA dataset for multilingual factual probing [14]. It comprises 37,498 instances across 43 relations, formatted as cloze prompts, *e.g.*, "[X] plays [Y] music," where subject, relation, object form a triplet.

Models: We use encoder-based ML-LMs, including multilingual BERT (mBERT) [1] and XLM-R [18] for knowledge probing. We focus on encoder-based models rather than generative ones because they are smaller yet still exhibit strong performance on language understanding tasks. For our factual knowledge probing task, which employs fill-in-the-blank queries, encoder-based models excel at integrating information across entire sentences, ensuring a detailed contextual understanding.

Evaluation protocol: We substitute X with the subject and replace Y with mask tokens in each relational template to form a query (*e.g.*, "The Beatles play [MASK] music.") and feed it to ML-LMs. If, in this instance, it predicts the mask token to be "rock," we consider that ML-LMs capture the fact. Since the object is not necessarily tokenized as a single token, we set the exact number of mask tokens corresponding to the object in the template and let ML-LMs predict multiple mask tokens simultaneously.

Results: Figure 2 shows the first-rank precision (P@1) across all languages with mBERT.[1)] We can observe low P@1 scores for most low-resource languages, and different languages largely differ from each other in recallable facts. As mBERT outperforms XLM-R in most languages, we will primarily focus on mBERT, a 12-layer Transformer-based ML-LM pre-trained on Wikipedia text across 103 languages for clarity in our subsequent analysis.

	it	ja	af
mBERT P@1	16.94	1.34	12.05
One-token P@1	15.27	15.34	17.00
One-token entities	1675	126	498
XLM-R P@1	10.80	4.78	8.17
One-token P@1	13.67	14.73	16.58
One-token entities	923	244	333

Table 1 P@1 and one-token object counts for mBERT and XLM-R in Italian (it), Japanese (ja) Afrikaans (af).

3 Factors Behind Probing Gaps

Figure 2 shows that factual probing accuracy for various languages exhibits substantial differences. In this section, we will evaluate the potential factors contributing to these differences and examine how they relate to the proficiency of ML-LMs in cross-lingual transfer.

Training data volume: The first factor relates to the amount of distinct factual knowledge seen during the training of ML-LMs We use the training data volume to estimate the amount of factual knowledge in the training data, specifically the data size of Wikipedia[2)] abstracts and full articles. Then, we calculate the Pearson correlation coefficient between probing accuracy (P@1) and data volumes, yielding values of 0.44 and 0.51 for abstracts and full articles, respectively. These moderate correlations suggest that training data volume has a limited impact on learning factual knowledge, implying that other factors contribute to the acquisition of facts by ML-LMs. The details of abstract size and P@1 are shown in Figure 2.

Number of mask tokens: There are correlations of −0.81 (mBERT) and −0.74 (XLM-R) between P@1 and the number of subwords in the target entities. As shown in Table 1,

1) Refer to Appendix A for language codes and detailed accuracies for both mBERT and XLM-R.

2) We use Wikipedia dumps prior to mBERT's release.

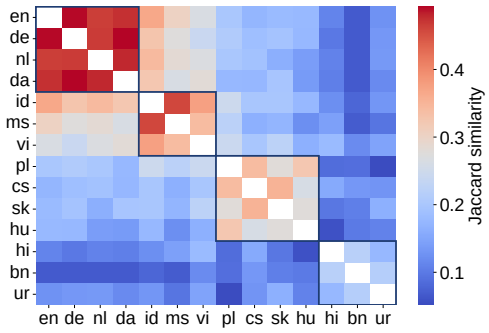

Figure 3 Jaccard similarity matrix of shared factual knowledge across languages with mBERT.

while both ML-LMs have similar P@1 scores for predicting one-token entities, XLM-R captures more one-token entities in Japanese (ja), resulting in more accurate predictions. However, the mask token and training data volume cannot fully explain the P@1 differences across languages, as Afrikaans (af) outperforms Japanese (ja) for one-token P@1 even with Japanese having ten times more training data than Afrikaans, as shown in Figure 2.

Localized knowledge cluster: We hypothesize that the high accuracy for low-resource languages may result from the model's proficiency in cross-lingual factual knowledge sharing. To investigate this, we assess shared facts between languages using Jaccard similarity. Figure 3 shows that languages in geographical proximity exhibit greater overlap in shared facts. Geographically close languages, such as Indonesian (id), Malay (ms), and Vietnamese (vi), demonstrate higher similarities, indicating substantial shared content. This suggests that cross-lingual knowledge transfer does not occur universally across all languages. Rather, it seems to be localized, influenced more by shared culture and vocabulary. We will explore this phenomenon further in the following sections.

4 Cross-lingual Representation

This section examines how ML-LMs represent facts within their parameter spaces through two scenarios. In the first scenario, facts are independently maintained in different languages (Figure 1(a)), which we refer to as "language-independent." In the second, fact representations are unified across languages in an embedding space (Figure 1(b,c)), called "cross-lingual" representations.

Factual neuron probing: Building on the theory that specific neurons in the feed-forward network (FFN) store facts [19, 20], we analyze the cross-lingual representation

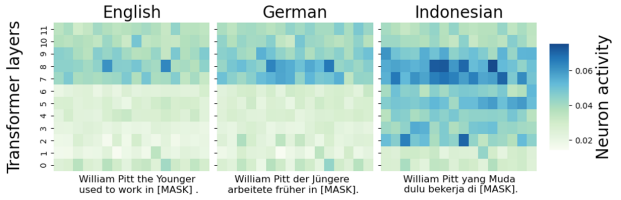

Figure 4 Neuron activities in mBERT for three languages, in response to an identical fact. Color intensity implies neuron activity, with neurons in each Transformer layer grouped into 16 bins. Distinct activation patterns in the English-Indonesian pair indicate cross-lingual representation.

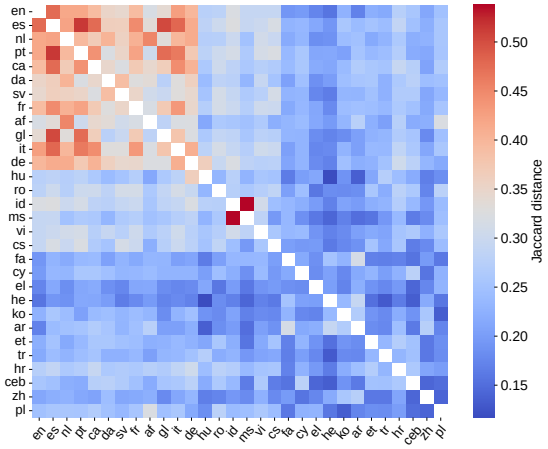

Figure 5 Language similarity based on top 50 shared active neurons by probing on mLAMA with mBERT.

of facts using PROBELESS [21], an efficient neuron attribution method that measures neuron importance in representing facts. Specifically, we collect the active neurons for the same fact in various languages to identify cross-lingual or language-independent fact representations. Languages with similar neuron activity patterns suggest a cross-lingual representation of that fact.

Do cross-lingual representations exist? Through a case study of neuron probing (Figure 4), we find that while some languages exhibit similar neuron activities for a given fact, others may exhibit distinct patterns, indicating the presence of both language-independent and cross-lingual representations. To measure the extent of cross-lingual sharing of a specific fact, we calculate the Jaccard similarity between the top 50 active neurons of two languages. We then compute pairwise language similarities by averaging the Jaccard similarity across all their shared facts, as shown in Figure 5. Figure 5 shows that there are no consistent geographical boundaries among languages, suggesting that both the language-independent scenario and the cross-lingual sharing scenario largely depend on specific facts.

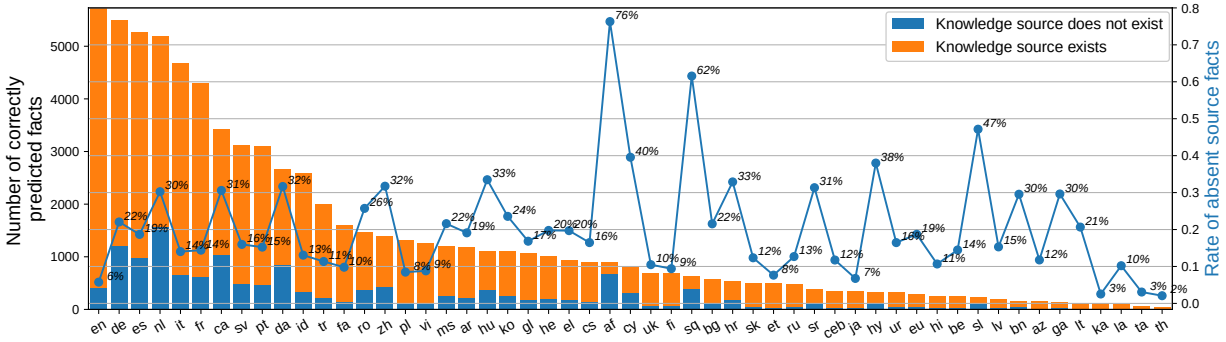

Figure 6 Number of correctly-predicted facts with mBERT in terms of the existence of knowledge source.

5 Cross-lingual Share vs. Transfer

We subsequently explore the formation of cross-lingual representations within ML-LMs to assess whether they are learned individually from distinct language corpora and subsequently aligned into a common semantic space (Figure 1(b)) or whether they are acquired through cross-lingual transfer (Figure 1(c)).

Tracing the roots of facts back to data: We use a simple yet effective method to check the presence of a fact in text: for a fact triplet (subject, relation, object), we examine the occurrences of the subject and object in mBERT's training data, Wikipedia. If both can be found, the fact is considered present in the data. Although this approach may not provide precise quantitative results, it is useful for exploring cross-lingual transfer possibilities. See § B for a detailed description of the method for checking subject/object occurrences. We assess the absence rates of all facts and correctly predicted facts, respectively. As shown by the results for 53 languages in Figure 6, languages with more training data exhibit better factual knowledge coverage, as expected. Nevertheless, several facts, such as those in Afrikaans (af) and Albanian (sq), are accurately predicted despite not having verifiable existence in the training corpus, suggesting a high possibility of cross-lingual transfer.

What kinds of facts are absent yet predictable? Analysis reveals that many of the facts that are absent in the knowledge source but correctly predicted were relatively easy to predict. We categorize these easy-to-predict facts into two types: **shared entity tokens** and **naming cues**. The former refers to queries in which the target object shares tokens with the subject (e.g., 'Sega Sports R&D is owned by Sega.'), while the latter pertains to entity-universal associations across person names, countries, and languages (e.g., 'The native language of Go Hyeon-jeong is Korean.'). In both cases, ML-LMs can predict the object entity from the subwords of the subject entity. However, some other facts are difficult to infer from the entities alone (e.g., 'Crime & Punishment originally aired on NBC'), suggesting a high possibility of cross-lingual transfer. We classify facts into the three types by rule-based method, as detailed in § C.

We measure the average proportions of facts correctly predicted by mBERT for the three types across languages: shared entity tokens (25.8%), naming cues (22.0%), and others (52.2%). The predictability of easy-to-predict facts suggests that ML-LMs can rely on simple deductions rather than encoding specific facts to make predictions, highlighting the need to enhance probing datasets to enable a more effective factual knowledge evaluation. Meanwhile, the high ratio of predictable facts that are not easy to predict suggests that ML-LMs indeed possess cross-lingual transfer ability for factual knowledge for some languages. Refer to § C for details.

6 Conclusions

Our research establishes the groundwork for further studies in understanding cross-lingual factual knowledge representation. Through comprehensive factual knowledge probing and analysis across 53 languages, we evaluate factors affecting cross-lingual knowledge transfer on factual knowledge, such as the training data volume and mask token count, and identify knowledge-sharing patterns among geographically close languages. We then leverage neuron probing and propose knowledge tracing methods to uncover three multilingual knowledge representation patterns in ML-LMs: language-independent, cross-lingual shared, and transferred. Our future work will investigate the knowledge representations in generative large LMs like Aya [2].

Acknowledgements

This work was partially supported by the special fund of Institute of Industrial Science, The University of Tokyo, by JSPS KAKENHI Grant Number JP21H03494, and by JST, CREST Grant Number JPMJCR19A, Japan.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[2] Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, et al. Aya model: An instruction finetuned open-access multilingual language model, 2024.

[3] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics.

[4] Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 12365–12394, Singapore, December 2023. Association for Computational Linguistics.

[5] Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. First align, then predict: Understanding the cross-lingual ability of multilingual BERT. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume**, pp. 2214–2231, Online, April 2021. Association for Computational Linguistics.

[6] Tyler Chang, Zhuowen Tu, and Benjamin Bergen. The geometry of multilingual language model representations. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 119–136, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[7] Tianze Hua, Tian Yun, and Ellie Pavlick. mOthello: When do cross-lingual representation alignment and cross-lingual transfer emerge in multilingual models? In Kevin Duh, Helena Gomez, and Steven Bethard, editors, **Findings of the Association for Computational Linguistics: NAACL 2024**, pp. 1585–1598, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[8] Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. Cross-lingual ability of multilingual bert: An empirical study. In **The Eighth International Conference on Learning Representations**, 2020.

[9] Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. Emerging cross-lingual structure in pretrained language models. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 6022–6034, Online, July 2020. Association for Computational Linguistics.

[10] Ibraheem Muhammad Moosa, Mahmud Elahi Akhter, and Ashfia Binte Habib. Does transliteration help multilingual language modeling? In **Findings of the Association for Computational Linguistics: EACL 2023**, pp. 670–685, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[11] Machel Reid and Mikel Artetxe. On the role of parallel data in cross-lingual transfer learning. In **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 5999–6006, Toronto, Canada, July 2023. Association for Computational Linguistics.

[12] Jiahuan Li, Shujian Huang, Aarron Ching, Xinyu Dai, and Jiajun Chen. PreAlign: Boosting cross-lingual transfer by early establishment of multilingual alignment. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 10246–10257, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[13] Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 5943–5959, Online, November 2020. Association for Computational Linguistics.

[14] Nora Kassner, Philipp Dufter, and Hinrich Schütze. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume**, pp. 3250–3258, Online, April 2021. Association for Computational Linguistics.

[15] Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. GeoMLAMA: Geo-diverse commonsense probing on multilingual pre-trained language models. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 2039–2055, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[16] Constanza Fierro and Anders Søgaard. Factual consistency of multilingual pretrained language models. In **Findings of the Association for Computational Linguistics: ACL 2022**, pp. 3046–3052, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[17] Amr Keleg and Walid Magdy. DLAMA: A framework for curating culturally diverse facts for probing the knowledge of pretrained language models. In **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 6245–6266, Toronto, Canada, July 2023. Association for Computational Linguistics.

[18] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics.

[19] Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. Analyzing individual neurons in pre-trained language models. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 4865–4880, Online, November 2020. Association for Computational Linguistics.

[20] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 8493–8502, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[21] Omer Antverg and Yonatan Belinkov. On the pitfalls of analyzing individual neurons in language models. In **International Conference on Learning Representations**, 2022.

ISO (Lang.)	mBERT	XLM-R	ISO (Lang.)	mBERT	XLM-R
en (English)	**19.07**	17.08	cs (Czech)	**5.63**	1.21
id (Indonesian)	**18.15**	13.99	ceb (Cebuano)	**5.11**	0.76
it (Italian)	**16.94**	10.80	et (Estonian)	**4.97**	3.82
de (German)	**16.91**	12.06	sq (Albanian)	**4.93**	3.31
es (Spanish)	**16.65**	10.51	sk (Slovak)	**4.90**	2.84
nl (Dutch)	**15.98**	10.47	bg (Bulgarian)	4.51	**5.07**
pt (Portuguese)	**14.76**	14.05	ur (Urdu)	**4.41**	4.40
ca (Catalan)	**14.11**	5.23	uk (Ukrainian)	**3.84**	0.64
tr (Turkish)	**14.08**	13.79	fi (Finnish)	3.58	**4.43**
da (Danish)	**13.56**	12.01	hy (Armenian)	3.25	**3.90**
ms (Malay)	**13.14**	11.20	sr (Serbian)	**3.07**	2.45
sv (Swedish)	**12.89**	11.63	hi (Hindi)	2.95	**3.78**
fr (French)	**12.68**	7.79	be (Belarusian)	**2.80**	0.78
af (Afrikaans)	**12.05**	8.17	eu (Basque)	**2.45**	1.19
ro (Romanian)	11.33	**13.38**	lv (Latvian)	**2.15**	1.66
vi (Vietnamese)	10.93	**11.78**	az (Azerbaijani)	**1.99**	3.21
gl (Galician)	**10.00**	6.04	ru (Russian)	**1.90**	0.79
fa (Persian)	**8.67**	7.30	bn (Bangla)	1.76	**2.67**
cy (Welsh)	**7.98**	5.08	ka (Georgian)	1.45	**1.89**
el (Greek)	**7.24**	5.68	ja (Japanese)	1.34	**4.78**
he (Hebrew)	**6.78**	4.60	sl (Slovenian)	1.26	**1.77**
ko (Korean)	**6.73**	7.18	lt (Lithuanian)	1.25	**2.31**
zh (Chinese)	**6.51**	4.05	la (Latin)	1.21	**1.83**
pl (Polish)	**6.33**	5.09	ga (Irish)	**0.96**	0.56
ar (Arabic)	6.11	**6.16**	ta (Tamil)	0.90	**0.93**
hu (Hungarian)	**5.86**	5.42	th (Thai)	0.49	**2.75**
hr (Croatian)	**5.65**	2.36	**Average (macro)**	**8.85**	6.88

Table 2 P@1 for 53 languages on mBERT with both mBERT and XLM-R, with all language codes.

A Full Probing Accuracies

Table 2 lists the probing P@1 for the 53 languages on mLAMA with mBERT and XLM-R, respectively, to complement the overall results.

B Occurrence Checking Method

We use subject-object co-occurrence as an approximation method to determine whether a fact is traced back to the data. We rigorously adhere to the preprocessing and sentence-splitting guidelines for mBERT [1]. Using the WikiExtractor,[3] we extract only text passages, deliberately omitting lists, tables, and headers. Each extracted document is segmented into multiple lines, with each line containing no more than 512 tokens.[4] Using string matching between the object-subject pair and Wikipedia text, we assess the co-occurrence of the object and subject for a given fact. If there is a co-occurrence, we consider the fact present; otherwise, it is considered absent.

3) https://github.com/attardi/wikiextractor

4) The maximum number of tokens that can be input to mBERT in training.

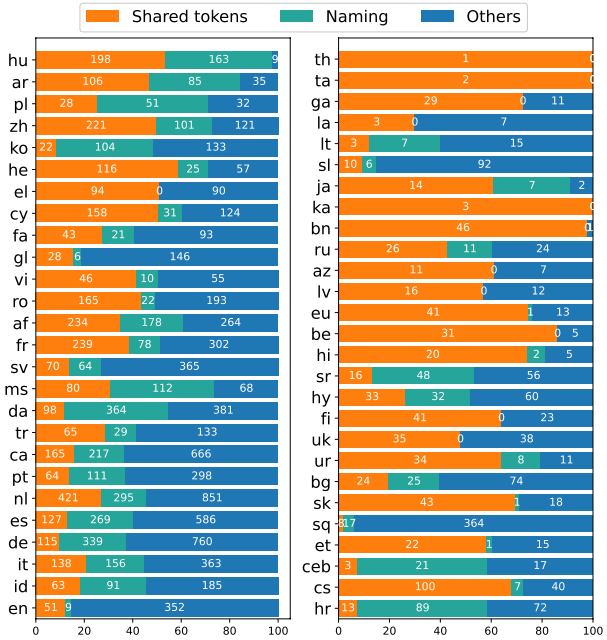

Figure 7 The count of three types of absent and predictable facts with mBERT.

C Classifying Predictable Facts

We classify the three types of predictable facts by the following rules.

Shared entity tokens: We normalize entities by rules, such as lowercasing strings and unifying Chinese traditional/simplified characters, and then assess if the object is a substring of or shares subwords with the subject.

Naming cues: We manually select several relations containing information among person name, location, and country entities.

Others: The facts other than those classified into shared entity tokens and naming cues are regarded as others.

Following the rules above, we classify the predictable facts in each language into these three types and measure their count, as shown in Figure 7. It shows that even without the Without easy-to-predict facts, the absence rate drops but is still not zero for some of the lan- guages (blue bar in Figure 7), such as Albanian (sq), Slovenian (sl), and Galician (gl), indicating that ML-LMs indeed possess cross-lingual transfer ability for factual knowledge for some languages.