Dataset	Method	m-BERT	m-GPT
English	LIKN (Ours)	2.359 (↑10.29%)	2.542 (↑5.92%)
	Mono-KN	2.139	2.400
	Seq-KN	3.800	4.285
Chinese	LIKN (Ours)	7.175 (↑213.05%)	8.868 (↑277.36%)
	Mono-KN	2.292	2.350
	Seq-KN	4.092 (↓42.97%)	3.654 (↓58.80%)

Table 2: Results of cross-lingual knowledge editing. LIKN represents editing language-independent knowledge neurons, Mono-KN denotes editing knowledge neurons in one language's dataset corresponding to another, and Seq-KN denotes sequentially editing knowledge neurons in two languages. The symbol '↑' shows a success rate increase in LIKN over Mono-KN, represented as $\frac{\text{LIKN}-\text{Mono-KN}}{\text{Mono-KN}}$, and '↓' indicates a decrease in LIKN compared to Seq-KN, represented as $\frac{\text{LIKN}-\text{Seq-KN}}{\text{LIKN}}$.

the changes in the corresponding facts in another language. (2) Sequentially editing the knowledge neurons of two languages, observing the changes in the corresponding facts in both languages.

Our analysis of Table 2 brings to light two insights.

(1) Language-independent knowledge neurons facilitate cross-lingual editing. Compared to editing in Chinese or English, editing language-independent knowledge neurons has a higher success rate in all settings; in the Chinese dataset, the success rates for m-BERT and m-GPT increased by 213.05% and 277.36%. This indicates that while editing facts in one language and expecting changes in another is challenging, language-independent neurons provide a viable solution.

(2) Editing each language separately does not guarantee better results. Though one might intuitively edit each language to achieve cross-lingual changes, our experiments show that this method not only relies on more computational resources but also might underperform. Sequential editing led to 42.97% and 58.80% lower success rates for m-BERT and m-GPT respectively, compared to using language-independent neurons, possibly due to confusion from multiple edits. This emphasizes the importance of language-independent neurons.

3.4 Degenerate Knowledge Neurons and Fact-Checking Experiment

noindent**Identification of Degenerate Knowledge Neurons in Multilingual PLMs** We set up an experiment using module 3 to investigate the degenerate knowledge neurons, and the results are displayed in Figure 4. From our observations, degenerate knowledge neurons in m-BERT and m-GPT exhibit distribution patterns similar to knowledge neurons. This not only demonstrates a strong correlation between the degeneracy of factual knowledge and the facts themselves, but also reflects the PLMs' mastery of the facts.

noindent**Identification of degenerate knowledge neurons in Monolingual PLMs** In our experiments with monolingual PLMs, we successfully identify the degenerate knowledge neurons and prove that they are inherently present within the PLMs. A possible question regarding degenerate knowledge neurons is: does the PLMs store the same fact in multiple languages, thus utilizing multiple neuron sets for the same information? To dispel this notion and demonstrate that the existence of degenerate knowledge neurons is unrelated to the support of multilingualism in the PLMs, we extend our exploration to monolingual PLMs, specifically in BERT and GPT-2. The distributions of these degenerate knowledge neurons is depicted in Figure 5, further reinforcing our conclusion.

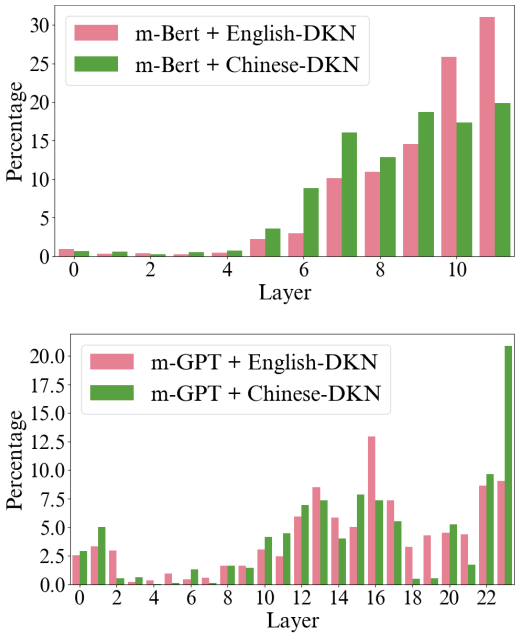

Figure 4: The distributions of degenerate knowledge neurons (DKN) in multilingual PLMs under two languages.

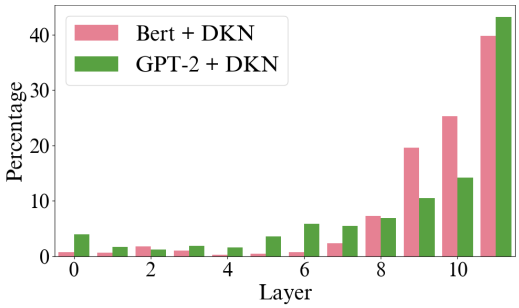

Figure 5: The distributions of degenerate knowledge neurons (DKN) in monolingual PLMs under two languages.

noindent**Fact-Checking Experimental Settings and Results** PLMs may conceal false facts (**??**), and current solutions often rely on external data for fact-checking (**?**). Considering the nature of the functional overlap of degenerate knowledge neurons, we design a fact-checking experiment to detect wrong facts based on degenerate knowledge neurons without relying on external data. Next, we introduce our experimental settings in detail.

First, the mLAMA dataset is modified to include a *wrong*

fact attribute. For a fact triple associated with a certain relation name of fact, such as $\langle \textit{Tanzania}, \textit{Capital}, \textit{Dodoma} \rangle$, we randomly select an object (e.g., *Dar es Salaam*) from the same relation name as a wrong fact. Then, to validate the practical implications of our findings, we divide each type of query in the dataset into two parts proportionally. For each type, the first segment is used to obtain degenerate knowledge neurons, and we identify those exceeding a certain threshold of $t\%$ in quantity. Then, we take the queries from the second part, along with the corresponding correct or incorrect facts, as input and compute the average activation score of the degenerate knowledge neurons. If the average activation score surpasses a threshold λ, the fact is classified as correct. We use the original PLMs to directly evaluate the correctness of facts for comparative analysis. This configuration prevents the PLMs from employing the degenerate knowledge neurons of the query itself for fact-checking, rendering the experiments more convincing. We denote our method as "with_DKN" in the Table 3. Finally, since the current fact-checking method must rely on external data, we use the PLMs to directly perform fact-checking as the baseline of our method, denoted as "wo_DKN" in the Table 3. We use Precision, Recall and F1-score as evaluation metrics. The results in Table 3 lead us to the following conclusions.

(1) Degenerate knowledge neurons can help the PLMs detect wrong facts. Under various settings, our method is better than the baseline method, especially for Chinese datasets and auto-regressive models. For instance, in the context of m-GPT and Chinese datasets, the F1 score of our method has increased by 167150% compared to the baseline.This substantial improvement indicates that the presence of degenerate knowledge neurons enhances the PLMs' stable mastery of factual knowledge. (2) Using PLMs for fact-checking, they often judge a fact as correct, leading to extremely high Recall. This aligns with observations that generative language models may produce incorrect information if presented with a false premise (**???**). It is essential to recognize that a model's low predictive probability does not hinder the accurate identification of knowledge neurons. As shown in Equation 1, using the true value y^* allows for correct knowledge neuron localization even when the model's output is erroneous.

(3) Auto-regressive models show higher Recall than auto-encoding models. This may be due to the auto-regressive design favoring coherence over accuracy, and the auto-encoding possibly being more conservative (**?**).

(4) The existence of degenerate knowledge neurons is unrelated to the support of multilingualism in the PLMs. In the monolingual PLMs, i.e., BERT and GPT-2, fact-checking can also be performed based on degenerate knowledge neurons. This result further proves the existence of degenerate knowledge neurons and its usefulness.

Dataset	Model	Method	P	R	F1
English	m-BERT	wo_DKN	0.222	0.986	0.362
		with_DKN (Ours)	0.493	0.599	**0.541** (↑ 49%)
	m-GPT	wo_DKN	0.010	1.000	0.021
		with_DKN (Ours)	0.311	0.709	**0.433** (↑ 1962%)
Chinese	m-BERT	wo_DKN	0.010	1.000	0.020
		with_DKN (Ours)	0.870	0.524	**0.654** (↑ 3170%)
	m-GPT	wo_DKN	0.0002	1.000	0.0004
		with_DKN (Ours)	0.966	0.511	**0.669** (↑ 167150%)
English	BERT	wo_DKN	0.301	0.983	0.460
		with_DKN (Ours)	0.504	0.571	**0.535** (↑ 16%)
	GPT-2	wo_DKN	0.010	1.000	0.021
		with_DKN (Ours)	0.315	0.608	**0.415** (↑ 1876%)

Table 3: Fact-checking experiment results comparing methods with (with_DKN) and without (wo_DKN) degenerate knowledge neurons. The symbol "↑" shows F1-score improvement in with_DKN over wo_DKN as $\frac{\text{with\_DKN} - \text{wo\_DKN}}{\text{wo\_DKN}}$, with bold indicating the higher score.

4 Related Work

noindent**Knowledge Localization** Existing methods roughly fall into two categories: (1) Gradient-based method: **?**(**?**) first introduces the concept of knowledge neurons and localizes them by assessing the contribution of each neuron (**?**) through calculating their attribution scores using integrated gradients. (2) Causal-inspired method, introduced by **?**(**?**), defines knowledge neurons as the neuron activations within PLMs that have the strongest causal effect on predicting certain factual knowledge, and this method has inspired the creation of knowledge editing algorithms such as ROME (**?**), MEMIT (**?**), and MEND (**?**). However, current methods lack a universal approach for different PLM architectures and exploration in multiple languages.

noindent**Axiomatic Attribution Methods** **?**(**?**) introduces the axiomatic attribution method, emphasizing Sensitivity and Implementation Invariance as the core axioms for attribution methods, leading to Integrated Gradients (IG). Subsequent research includes Discretized IG (**?**), which uses interpolation strategies for gradient accuracy; Sequential IG (**?**) designed for word importance evaluation; and Effective Shapley value along with Shapley IG, developed by **?**(**?**) to enhance efficiency and effectiveness. We improve the baseline vectors for IG to minimize their information content.

5 Conclusion

In this research, we explore factual knowledge localization in multilingual PLMs using our architecture-adapted multilingual integrated gradient method. We further design two modules, leading to two discoveries of language-independent knowledge neurons and degenerate knowledge neurons. The former affirms that a portion of the knowledge in multilingual PLMs exists in a form that transcends language, while the latter introduces a novel type of neuron which is similar to the degeneration phenomenon observed in biological systems, and these neurons can be used to detect incorrect facts.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (No. 2020AAA0106400), the National Natural Science Foundation of China (No. 61976211, 62176257). This work is also supported by the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No.XDA27020100), the Youth Innovation Promotion Association CAS, and Yunnan Provincial Major Science and

Temporibus cum tenetur veniam laudantium suscipit