fact attribute. For a fact triple associated with a certain relation name of fact, such as $\langle$*Tanzania*, *Capital*, *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.

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.

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.

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