# On the Pitfalls of Analyzing Individual Neurons in Language Models

Omer Antverg Yonatan Belinkov
omer.antverg@cs.technion.ac.il belinkov@technion.ac.il
Technion – Israel Institute of Technology

## Abstract

While many studies have shown that linguistic information is encoded in word representations, few have studied **individual neurons** in these representations, to show how and in which neurons (dimensions) it is encoded. Such knowledge can be used to control the model's output or prune the model, and it can give us parameter-level explanations of the model's decisions.

Among individual neurons studies, the common approach is to use an external probe to rank neurons according to their relevance to some linguistic attribute, and to evaluate the obtained ranking using the same probe that produced it. We show two pitfalls in this methodology:

- The ranking evaluation approach confounds distinct factors: **probe** quality and **ranking** quality. We show that this approach is flawed, as certain probes with an intentionally **bad** ranking can surpass others with a good ranking.
- It focuses on encoded information, rather than information that is used by the model. We show that these are not the same, by developing a new ranking evaluation approach that focuses on **causality** using **interventions**.

We compare two recent ranking methods and a simple one we introduce, and evaluate them with regard to both of these aspects.

## Rankings and Data

We work with three methods to rank neurons according to their importance for a morphological attribute:

- **Linear** [1]: Train a linear classifier on word representations to learn some task $F$. Then, use the trained classifier's weights to rank the neurons according to their importance for $F$.
- **Gaussian** [3]: Train a generative classifier on the task $F$, based on the assumption that each dimension in $\{1, ..., d\}$ is Gaussian-distributed. Then, greedily select the most informative neuron, according to the classifier's performance, at every iteration.
- **Probeless**: For every attribute label $z \in Z$, calculate $q(z)$, the mean vector of all representations of words that possess the attribute and the value $z$. Then, calculate the element-wise difference between the mean vectors,

$$r = \sum_{z,z' \in Z} |q(z) - q(z')| \tag{1}$$

and obtain a ranking by arg-sorting $r$, i.e., the first neuron in the ranking corresponds to the highest value in $r$.

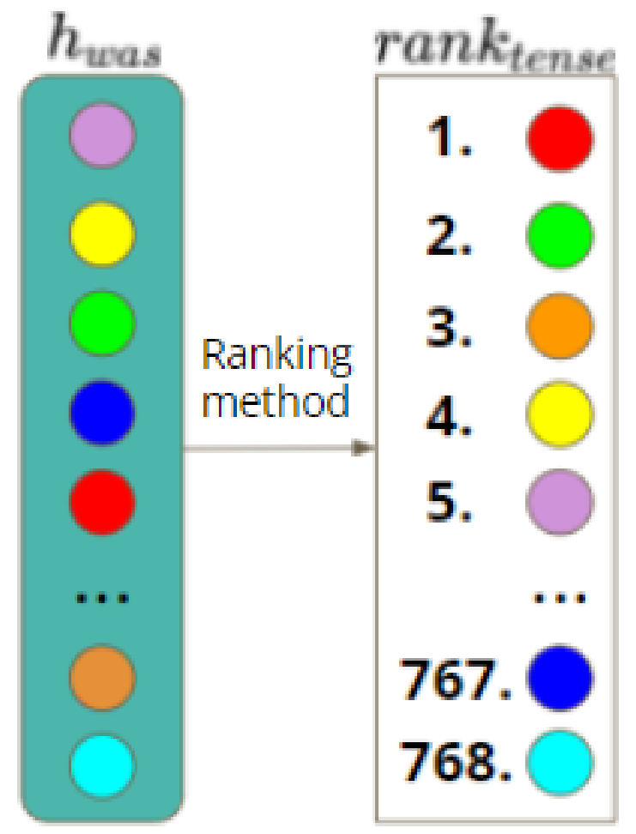


Figure 1. Neuron ranking Illustration. We rank the neurons of the representation of the word "was" according to their importance to some attribute, e.g., tense.

We use top-to-bottom and bottom-to-top versions of each of the rankings, as well as a random ranking baseline, ending up with 7 rankings overall.

### Data

- 9 different languages from the UD treebanks: Arabic, Bulgarian, English, Finnish, French, Hindi, Russian, Spanish and Turkish.
- Tasks: predictions of morphological attributes: Animacy, Aspect, Case, Definiteness, Gender, Mood, Number, Part of Speech, Person, Polarity, Possession, Tense, Voice.
- Representation: layers 2, 7 and 12 of a pre-trained M-BERT [2] model.
- Overall: 156 configs (language×attribute×layer).

## References

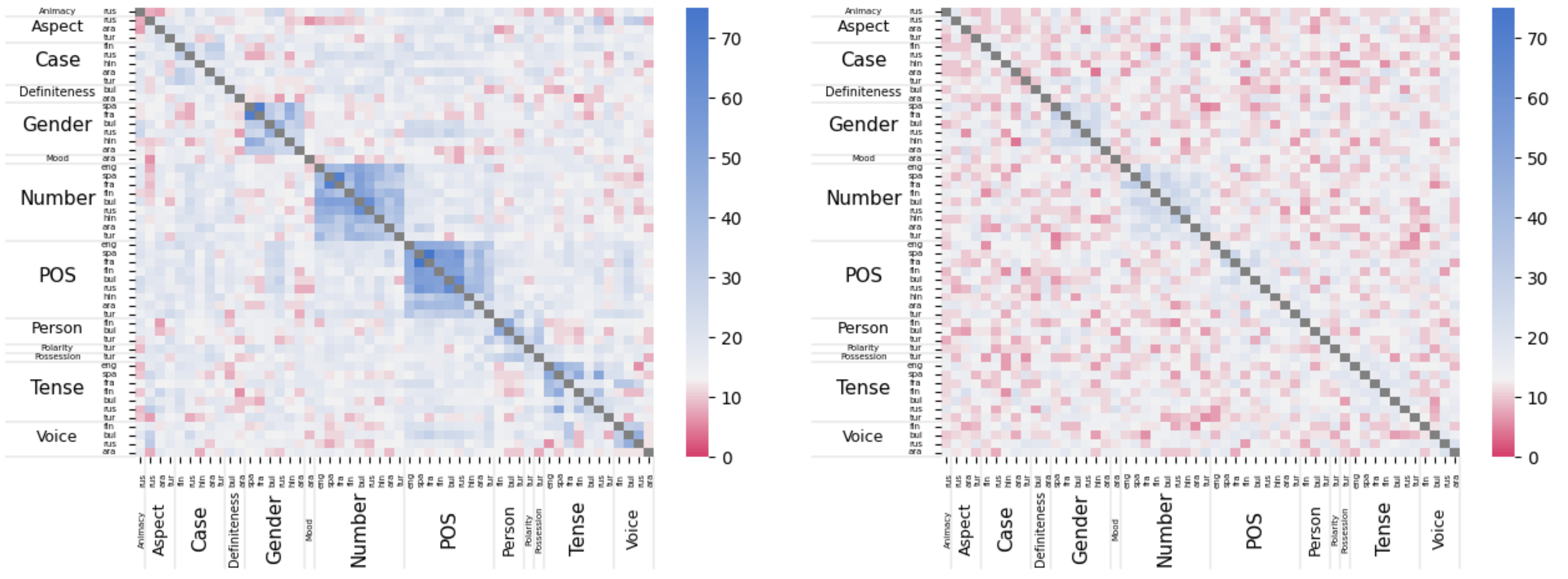
[1] Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James R. Glass. What is one grain of sand in the desert? analyzing individual neurons in deep NLP models. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, pages 6309–6317. AAAI Press, 2019.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[3] Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. Intrinsic probing through dimension selection. In *Proceedings of EMNLP*, pages 197–216, Online, 2020. Association for Computational Linguistics.

## Overlaps

- Multilingual models capture some universal concept of morphological attributes, independent of language.
- Significant overlaps between important neurons for the same attribute across languages, using Probeless.
- For Gaussian we do not see the same behaviour, meaning it is less consistent across languages.



(a) Layer 7 neurons overlap, using Probeless ranking. (b) Layer 7 neurons overlap, using Gaussian ranking.

## Ranking Evaluation by Probing

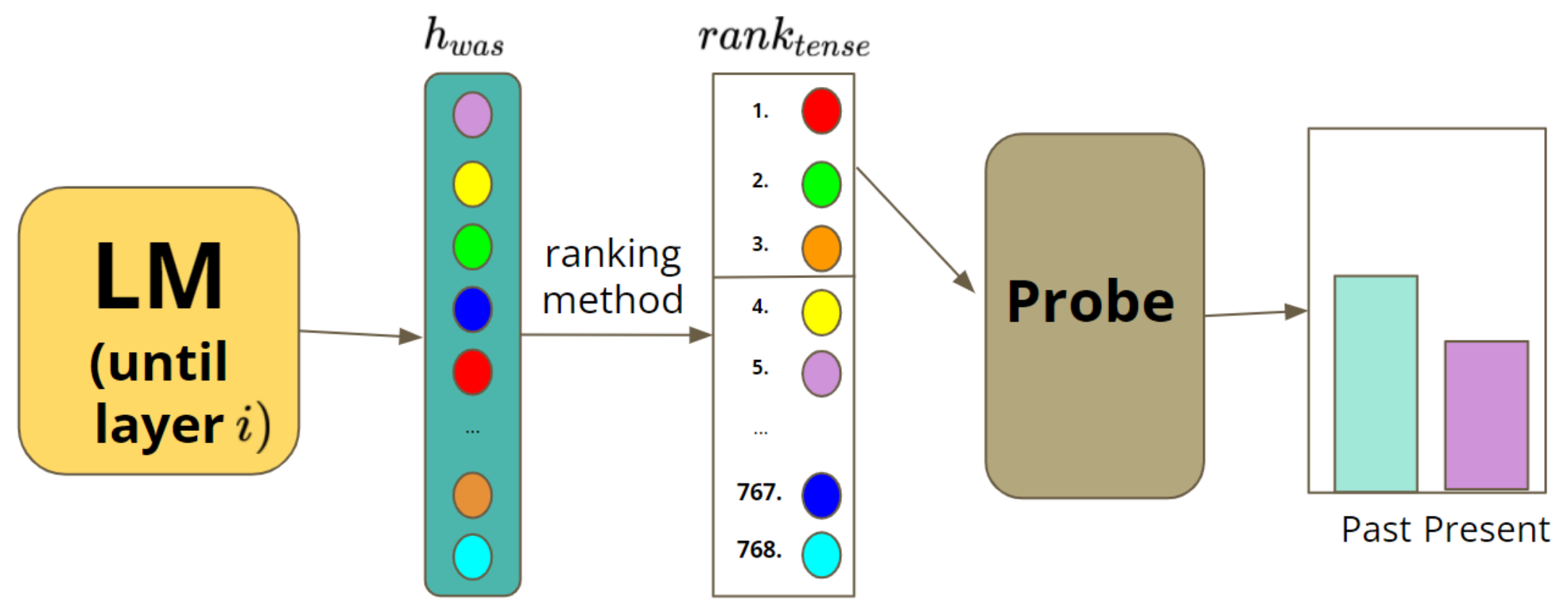
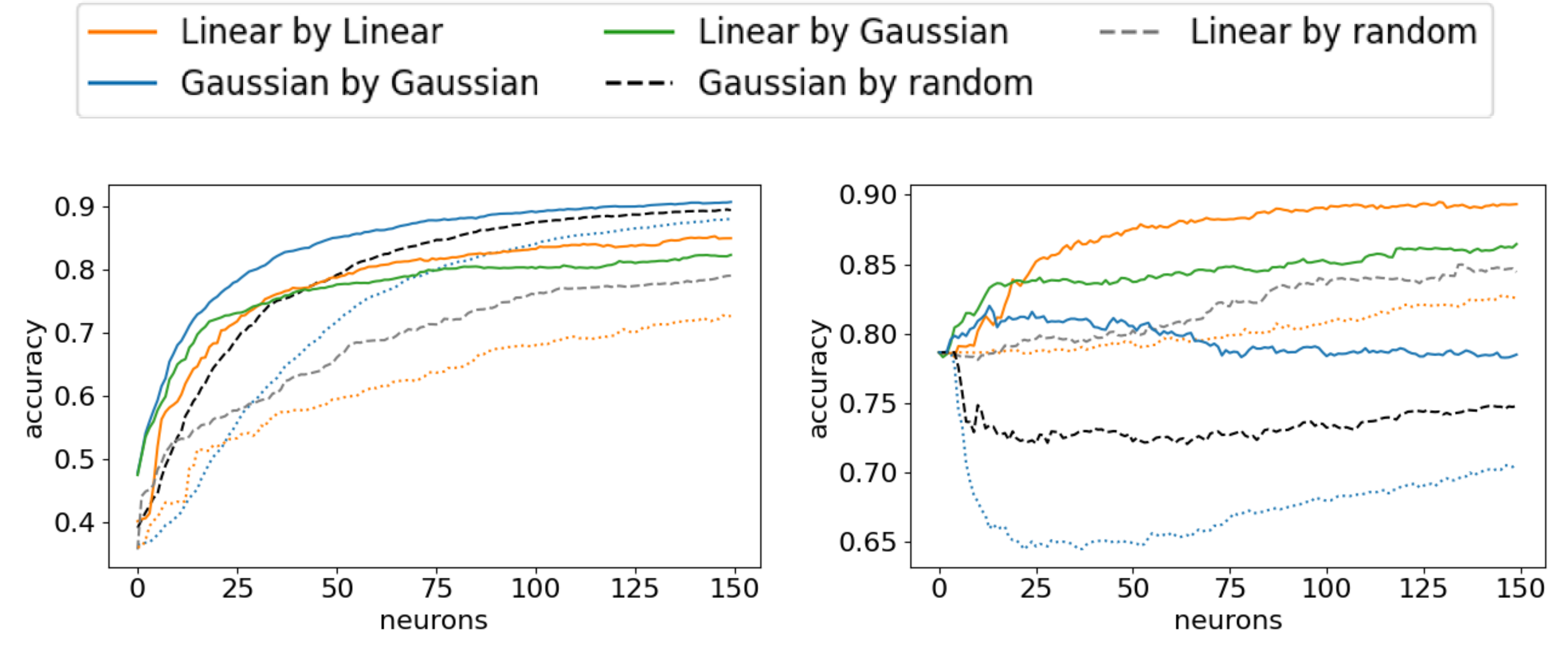


Figure 3. Ranking evaluation by probing: The probe is trained using only the $k$-highest ranked neurons.

On the left, the Gaussian probe using its **worst** ranking surpasses the Linear probe using its best ranking. On the right, it is the opposite: The Linear probe using its **worst** ranking surpasses the Gaussian probe using its best ranking. **These phenomena demonstrate that this evaluation approach is flawed.**



(a) Hindi part of speech layer 12 (G>L pattern). (b) Russian animacy layer 2 (L>G pattern).

Figure 4. Solid lines are top-to-bottom rankings; dashed are random rankings; dotted are bottom-to-top. Some lines are omitted for clarity.

## Ranking Evaluation by Interventions

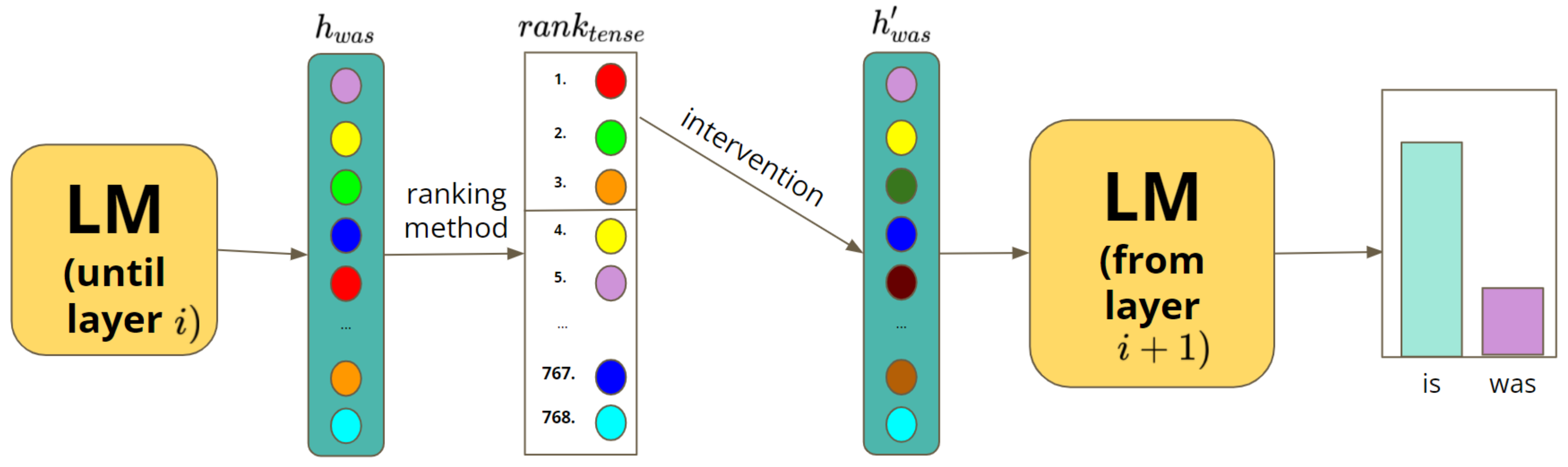


Figure 5. Ranking evaluation by interventions.

- For a representation $h$ with attribute label $z$ and ranking $\Pi(d)$, we modify the $k$-highest-ranked neurons by the ranking, $\Pi(d)_{[k]}$, by **subtracting the mean value of $z$ and adding the mean value of another label $z'$, multiplied by some factor.** Formally:

$$h_{\Pi(d)_{[k]}} = h_{\Pi(d)_{[k]}} + \alpha_k(q(z')_{\Pi(d)_{[k]}} - q(z)_{\Pi(d)_{[k]}}) \tag{2}$$

where $q$ is the same as in Probeless ranking, $z'$ is an attribute label such that $z \neq z'$, and $\alpha \in \mathbb{R}^d$ is a $\log$-scaled coefficients vector in the range $[0, \beta]$, such that the coefficient of the highest-ranked neuron is $\beta$ and that of the lowest-ranked neuron is $0$, and $\beta$ is a hyperparameter.
- **If the predicted word is different than the original one, then the model uses the information we modified.**
- We look for words which have the same lemma as the original one, but a different attribute value, and mark them as CLWV (Correct Lemma, Wrong Value). For example, **if the word "makes" becomes "made" when intervening for tense, then it counts as a CLWV, but if it becomes "make" or "prepared" it does not**.
- **Probeless provides the best CLWV in most configs**, and does so using a smaller number of neurons, in contrast to probing experiments. It implies that **encoded information and used information are not necessarily the same.**

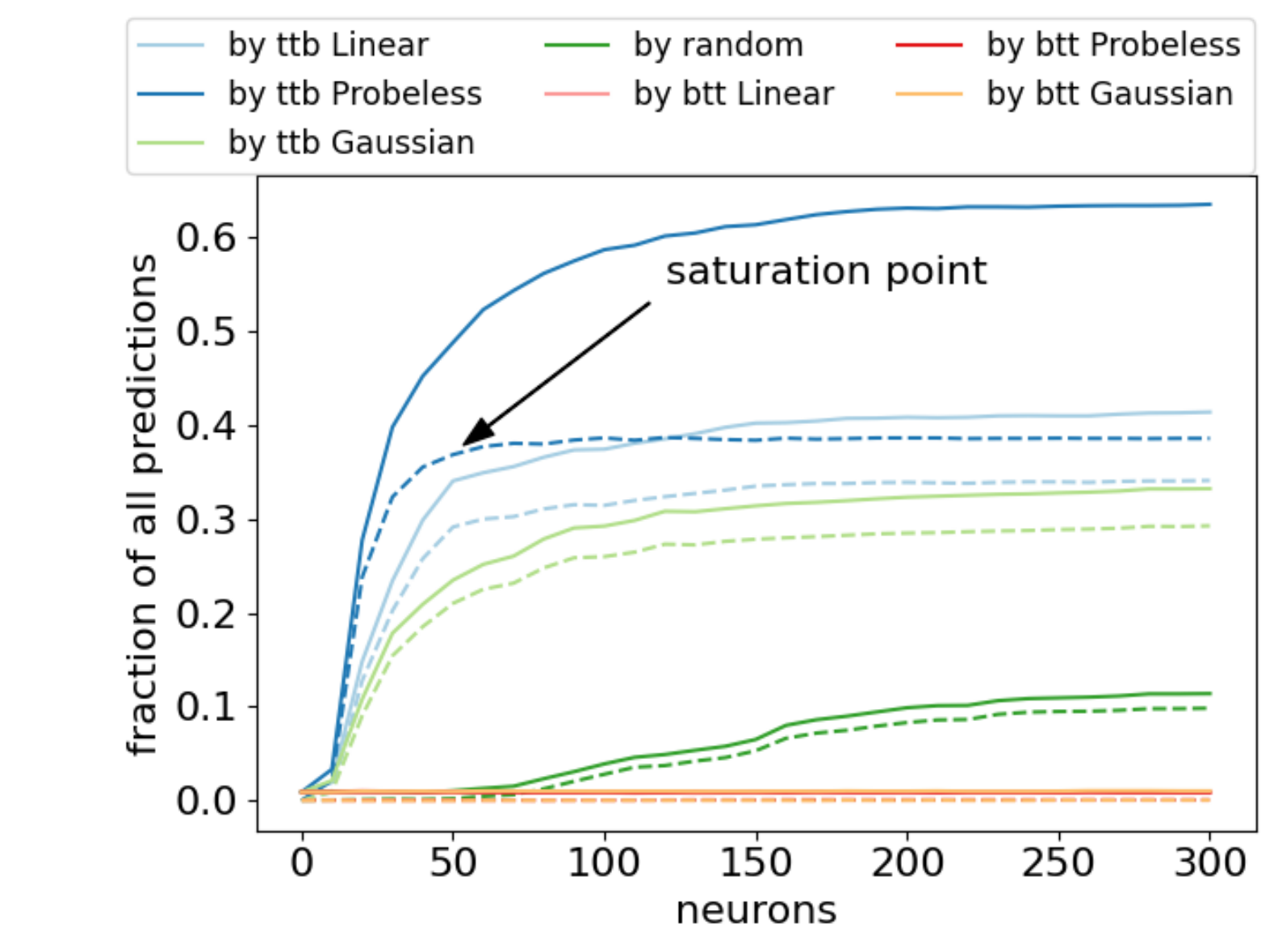


Figure 6. Spanish gender layer 2, intervention results with $\beta = 8$. Solid lines are error rates, dashed are CLWVs.

| | Linear | Gaussian | Probeless |
|---|---|---|---|
| English tense | 0.39, 60 | 0.26, 150 | 0.38, 30 |
| | 0.37, 50 | 0.34, 70 | 0.34, 30 |
| | 0.51, 60 | 0.41, 120 | 0.46, 30 |
| Spanish number | 0.28, 110 | 0.19, 100 | 0.35, 60 |
| | 0.26, 50 | 0.20, 40 | 0.25, 30 |
| | 0.23, 150 | 0.16, 140 | 0.40, 80 |
| Spanish gender | 0.29, 50 | 0.25, 80 | 0.37, 50 |
| | 0.29, 50 | 0.31, 50 | 0.33, 30 |
| | 0.26, 130 | 0.16, 110 | 0.35, 60 |

Table 1. CLWV value at saturation point and number of neurons modified at the saturation point, using the translation method.

## Conclusion

- We show that **previous ranking evaluation approach is flawed**, as the probe quality can heavily affect results.
- **We propose a new ranking evaluation approach**, that focuses on causality using interventions in the representations space.
- We present a new ranking method: Probeless, which is simple and fast to obtain, and **overcomes other rankings methods in causal evaluation.**
- We observe that **multilingual models capture some universal concept of certain morphological attributes, independent of language.**