Qualifying Knowledge and Knowledge Sharing in Multilingual Models

Nicolas Guerin *nicolas.p.guerin@gmail.com*
Laboratoire de Sciences Cognitives et Psycholinguistique
Dept d'Etudes Cognitives, ENS, PSL University, EHESS, CNRS

Ryan Nefdt *Ryan.Nefdt@uct.ac.za*
University of Cape Town
University of Bristol

Emmanuel Chemla *emmanuel.chemla@ens.psl.eu*
Laboratoire de Sciences Cognitives et Psycholinguistique
Dept d'Etudes Cognitives, ENS, PSL University, EHESS, CNRS
Earth Species Project

Reviewed on OpenReview: *https://openreview.net/forum?id=hnpB3SRbZj*

Abstract

Pre-trained language models (PLMs) have demonstrated a remarkable ability to encode factual knowledge. However, the mechanisms underlying how this knowledge is stored and retrieved remain poorly understood, with important implications for AI interpretability and safety. In this paper, we disentangle the multifaceted nature of knowledge: successfully completing a knowledge retrieval task (e.g., "*The capital of France is* ___") involves mastering underlying concepts (e.g., *France, Paris*), relationships between these concepts (e.g., *capital of*) and the structure of prompts, including the language of the query. We propose to disentangle these distinct aspects of knowledge and apply this typology to offer a critical view of neuron-level knowledge attribution techniques. For concreteness, we focus on Dai et al.'s (2022) Knowledge Neurons (KNs) across multiple PLMs (BERT, OPT, Llama and Gemma), testing 10 natural languages and additional unnatural languages (e.g. Autoprompt). Our key contributions are twofold: (i) we show that KNs come in different flavors, some indeed encoding entity level concepts, some having a much less transparent, more polysemantic role , and (ii) we address the problem of cross-linguistic knowledge sharing at the neuron level, more specifically we uncover an unprecedented overlap in KNs across up to all of the 10 languages we tested, pointing to the existence of a partially unified, language-agnostic retrieval system. To do so, we introduce and release the Multi-ParaRel dataset, an extension of ParaRel, featuring prompts and paraphrases for cloze-style knowledge retrieval tasks in parallel over 10 languages.

1 Introduction

Recent advances in Large Language Models (LLMs) have led to models trained on vast and diverse linguistic datasets drawn from across the Internet, incorporating numerous languages simultaneously (Scao et al., 2023; Touvron et al., 2023; Achiam et al., 2024). However, these languages are not evenly represented, and performance on low-resource languages often depends on cross-linguistic transfer from high-resource languages (Pires et al., 2019; Lample & Conneau, 2019; Conneau et al., 2020a; Huang et al., 2021). Whether LLMs can develop common, language-agnostic representations that enable such zero-shot transfer remains an open question in the literature (Singh et al., 2019; Kudugunta et al., 2019; Kassner et al., 2021). Kervadec et al.

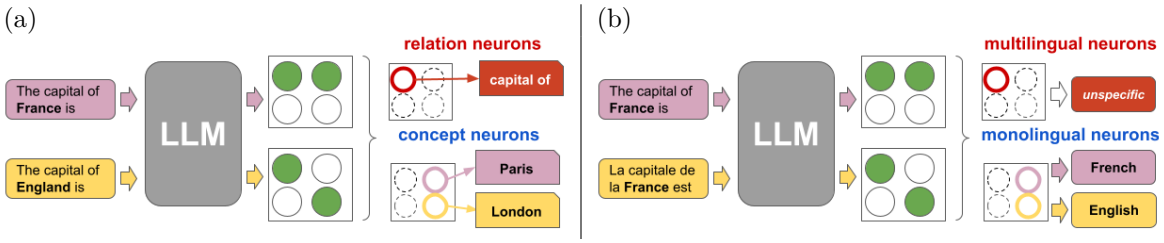


Figure 1: The Knowledge Neurons (KNs) hypothesis connects LLM success on a fill-in-the-blank cloze task (e.g. *The capital of France is*) to the activation of a small set of neurons. (a) The same neurons can be selected (green) in response to a single task, thereby qualifying as *concept* neurons (about e.g., Paris) or in response to a range of tasks all concerning a certain relations between concepts, thereby qualifying as *relational neurons* (e.g., *capital of* is a relation between France and Paris, between England and London, etc.). (b) In multilingual LLMs, concept and relational neurons may be selected specifically for a language or across languages.

(2023) extended this investigation to machine-generated languages, revealing that different representations can emerge, suggesting multiple ways knowledge may be encoded in LLMs.

Understanding how Pre-trained Language Models (PLMs) store and retrieve knowledge is essential for enhancing interpretability and safety in AI systems. Many recent studies have sought to localize and attribute specific knowledge to individual neurons within these models (Dai et al., 2022; Meng et al., 2022; 2023). These methods often attempt to identify neurons whose activations are critical for making accurate predictions. Typically, they focus on neurons in intermediate layers of Feed-Forward Networks (FFNs) within transformer architectures (Geva et al., 2021). These approaches face strong limitations, as highlighted in recent critiques (Hase et al., 2023; Niu et al., 2023; Huang et al., 2023).

In this work, we offer a novel perspective by refining the concept of "knowledge" itself. To correctly complete a prompt like *The capital of France is*, a model must process multiple layers of information: sensitivity to the specific concept *France*, retrieval of the target concept *Paris*, and understanding the relational context *capital of*. We introduce a method to distinguish these subtypes of knowledge—conceptual and relational—that is compatible with any knowledge attribution technique. We apply this method to the Knowledge Neurons (KNs) framework introduced by Dai et al. (2022), to provide a critical view on such a method and extend it to investigate how knowledge is shared across languages in PLMs (Figure 1).

Our contributions are:

- We propose a finer-grained typology of knowledge, providing a critical perspective on neuron-level attribution methods like the Knowledge Neuron hypothesis, in particular its expectation of monosemanticity.
- We analyze through this prism multiple PLMs (BERT, mBERT, OPT, Llama 2, and Gemma 2), revealing that a substantial number of 'Knowledge Neurons' exhibit polysemantic behavior, while others are specifically responsive to individual concepts or relations.
- We release `Multi-ParaRel`, a multilingual version of the `ParaRel` dataset (Elazar et al., 2021a), which includes 10 languages and is compatible with autoregressive models. The dataset is available here[1].
- We examine the extent to which knowledge representations are shared across languages at the level of individual neural units and demonstrate that LLMs store knowledge in similar neurons across 10 languages, and even in machine-generated languages (AutoPrompt), suggesting a shared cross-linguistic mechanism for knowledge retrieval.

[1]The dataset is available at `https://github.com/GpNico/multi-pararel`

2 Related Work

Multilingual Language Models The paradigm of Neural Machine Translation (NMT) has undergone a significant shift. Traditional approaches that relied on parallel corpora—or even synthetic data built from unaligned monolingual corpora (Sennrich et al., 2016; Lample et al., 2018; Artetxe et al., 2018)—are no longer the dominant standard. Instead, recent Large Language Models (LLMs) (Radford et al., 2019; Brown et al., 2020; Touvron et al., 2023; Achiam et al., 2024), trained on massive multilingual corpora scraped from the web, have become de facto multilingual systems. These models demonstrate strong cross-linguistic capabilities across a wide range of tasks (Devlin et al., 2019; Lample & Conneau, 2019; Conneau et al., 2020a; Liu et al., 2020; Xue et al., 2021; Scao et al., 2023; Vilar et al., 2023; Peng et al., 2023; Hendy et al., 2023; Bawden & Yvon, 2023).

Understanding how multilingual LLMs acquire and represent linguistic knowledge is critical to identifying their limitations and potential risks (Garcia et al., 2021; Raunak et al., 2021; Akhbardeh et al., 2021; Bapna et al., 2022). For instance, Guerreiro et al. (2023) explores how hallucinations affect translation quality in such systems. Yuemei et al. (2024) provides a comprehensive survey of multilingual LLMs (MLLMs), highlighting prevalent biases.

One core challenge is understanding how cross-linguistic capabilities emerge. Prior work has shown evidence of shared multilingual knowledge within models (Aharoni et al., 2019; Arivazhagan et al., 2019; Conneau et al., 2020b; K et al., 2020; Ri & Tsuruoka, 2022; Deshpande et al., 2022; Liu & Niehues, 2022; Choenni et al., 2023; Rajaee & Monz, 2024; Chua et al., 2025), though some findings are mixed. For example, Kudugunta et al. (2019) uses Singular Value Canonical Correlation Analysis to show that language representations in a NMT model are similar—particularly among related languages. Meanwhile, Singh et al. (2019) notes that mBERT tends to cluster representations by language, suggesting language separation despite shared architecture.

This is the context for our work. While shared cross-linguistic knowledge has been observed, we push this further by examining it at the neuron level. Most relevant to our approach, Chen et al. (2024) recently analyzed neuron overlap between English and Chinese. We extend this line of inquiry by comparing neuron activations across ten natural languages simultaneously. This broader comparison is essential. Pairwise analyses leave open the question of whether shared neurons generalize beyond specific language pairs. Overlap between two languages—especially if one is dominant (e.g., English)—might result from coincidental or biased patterns. But consistent sharing across ten diverse languages indicates a more robust, symmetrical multilingual encoding. Finally, we introduce a comparison with an 'unnatural language' (Shin et al., 2020), testing whether access to knowledge can be decoupled from linguistic form. These prompts are not human-interpretable and have been shown to elicit qualitatively different processing in LLMs (Kervadec et al., 2023), making them a stress test for the generality of shared representations. In fact, such an inclusive approach to training data is now even vindicated beyond languages, using non-linguistic sounds for training in speech (see Poli et al. (2024)), or even using several species at once for bioacoustics (e.g., Robinson et al. (2025)).

Knowledge in LLMs LLMs acquire knowledge by training on extensive corpora (Petroni et al., 2019; Roberts et al., 2020; Safavi & Koutra, 2021). The work by Petroni et al. (2019) introduced LAMA, a dataset designed to evaluate BERT through a fill-in-the-blank cloze task (e.g., *The capital of France is [MASK].*). Subsequent research has built upon LAMA (Jiang et al., 2021), highlighting the limitations of LLMs as knowledge bases (Elazar et al., 2021b; AlKhamissi et al., 2022; Wang et al., 2023; 2024b), while also attempting to enhance their performance (Wei et al., 2021; Petroni et al., 2020). Consequently, research has emerged focusing on localizing and editing knowledge directly within the model (Radford et al., 2017; Lakretz et al., 2019; Bau et al., 2020b; Sinitsin et al., 2020; Mitchell et al., 2021; 2022; De Cao et al., 2021; Santurkar et al., 2021; De Cao et al., 2022; Bau et al., 2020a; Cohen et al., 2023).

In this context, knowledge attribution methods such as ROME (Meng et al., 2022) and MEMIT (Meng et al., 2023) (both employing causal mediation techniques; Vig et al., 2020), along with Knowledge Neurons (Dai et al., 2022) (utilizing an integrated gradient approach; Sundararajan et al., 2017), have been proposed. These methods are predicated on the assumption that neurons within the intermediate layers of transformers' Feed-Forward Networks (FFNs) encode knowledge. However, we align with other studies (Hase et al., 2023; Niu

et al., 2023; Huang et al., 2023; Chen et al., 2025) that suggest this assumption may be an oversimplification. While certain neurons play a significant role in specific tasks (Lakretz et al., 2019; Manning et al., 2020; Rogers et al., 2020; He et al., 2024), LLM neurons are not necessarily monosemantic; rather, they can serve multiple functions depending on the context and task (Adly et al., 2024). Furthermore, their effectiveness in altering knowledge is subjective and widely debated (Hase et al., 2023). Other works (Wang et al., 2024a; Tang et al., 2024; Kojima et al., 2024) have identified multilingual neurons in LLMs; this paper focuses specifically on knowledge-related neurons, offering a more precise analysis. We propose a knowledge-attribution method-agnostic typology, illustrated with Dai et al.'s (2022) Knowledge Neurons. This approach aims to provide a critical view on the Knowledge Neurons hypothesis while exploring what insights it can offer regarding how knowledge is encoded in LLMs.

3 Methodological background

Knowledge The TREx dataset (Elsahar et al., 2018) is a collection of relational facts stored in triplets of the form $< h, r, t >$, with r a relation and h and t entities entering in that relation. TREx exhibit 41 relations, such as *being the capital of*, *was born in*, etc. Each full triplet can be referred to as an **instantiation** of its own relation r.

Knowledge Localization Methods Geva et al. (2021) observed that a FFN can be seen as a Key-Value memory system, similar to self-attention. To assess if and where knowledge could be stored in FFNs, Dai et al. (2022) used a knowledge attribution method based on integrated gradients (see next paragraph for details). They show that a fact (e.g., *The capital of France is Paris*) can be associated to a few neurons (around 4), whose activations correlate with the probability of the model to fill in the elements of the fact appropriately. Similarly, Meng et al. (2022) proposed Rank-One Model Editing (ROME), which uses causal mediation to localize and edit knowledge in GPT, and Meng et al. (2023) introduced Mass-Editing Memory in a Transformer (MEMIT), which edits facts at scale. All of these knowledge attribution methods have their limitations; we apply our analysis to the Knowledge Neurons by way of illustration. Our approach is applicable to all such methods.

Knowledge Neurons Dai et al. (2022) track Knowledge Neurons (KNs) during a fill-in-the-blank cloze task (see also Petroni et al., 2019) based on TREx. Let $w_i^{(l)}$ be the ith neuron of the intermediate layer of the lth FFN. The knowledge score of a neuron $w_i^{(l)}$ is calculated through the integrated gradient attribution method (Sundararajan et al., 2017), KNs are then filtered through thresholds. First, they retain only neurons with an attribution score greater than $t_{kn} \times \max_{i,l} \mathrm{Attr}_{h,p_r,t}(w_i^{(l)})$.This procedure is carried out for each prompt associated with a fact $< h, r, t >$, and thus yields a set of candidate KNs per prompt. Let us denote N_r the number of prompts for a given relation r. To get results robust to noise, and to factor out signal associated to specific prompts rather than knowledge, they keep only neurons appearing in the candidate neurons set of at least $p_{kn} \times N_r$ prompts. They propose thresholds of $t_{kn} = 0.2$ (only keep neurons scoring at least at 20% of the max attribution score) and $p_{kn} = 0.7$ (only keep neurons appearing in at least 70% of the different prompts for a given relation).

4 Method

Datasets For relational facts, we used the TREx dataset (Elsahar et al., 2018), which comprises 41 relations with approximately 1,000 facts per relation. For prompts, we employed the augmented version of `ParaRel` provided by Kervadec et al. (2023). This version retains only prompts compatible with autoregressive models and enriches the dataset with multiple paraphrases for each relation. In Section 6, we explore multilingual models, which we tested on the multilingual variant of LAMA (Kassner et al., 2021) as well as on a new multilingual version of `ParaRel` that we introduce. We refer to this new dataset as `Multi-ParaRel`.

The detailed methodology for creating `Multi-ParaRel`, along with a quality assessment, is provided in Appendix A. Our dataset currently spans 10 languages: English, French, Spanish, Catalan, Danish, German, Italian, Dutch, Portuguese, and Swedish. We also investigate an unnatural language: AutoPrompt. Following

the same train, development, and test splits as Shin et al. (2020), we trained 10 different seeds of AutoPrompt for each relation and each model. We also make these sets of prompts available.

Concept Neurons and Relation Neurons We propose a simple typology that refines the type of knowledge attributed while answering fill-in-the-blank cloze tasks. For example, correctly answering the question *What is the capital of France?* not only requires knowledge of the answer *Paris*, but also an understanding of the relationship between *France* and *Paris*. We thus introduce a simple principle: a neuron that is hypothesized to encode a specific concept, such as one about *Paris*, should not be also responsible for encoding other concepts, and should therefore not be associated to other facts such as *The capital of Spain is Madrid*. If a neuron consistently encodes the same relation across multiple instances, we refer to it as a relational neuron, indicating that it is sensitive to a relation, such as *capital of*.

We thus define **Relation Neurons** as KNs that appear in at least $t_r \times N$ instances of facts associated with a particular relation, where N is the total number of facts, and t_r is a predefined relational threshold. In contrast, neurons that appear in less than $t_c \times N$ of the facts, for some other threshold t_c, are referred to as **Concept Neurons**, as they are more likely to encode specific pieces of knowledge or information about individual entities.

The aim is to test the robustness of this distinction by investigating the role of the thresholds t_r and t_c. A 'clean' scenario that supports the Knowledge Neuron hypothesis and the idea of monosemanticity would show that some concept neurons are found even for $t_C \times N = 1$ (very specific to a concept), and relational neurons are found when $t_R \times N = N$ (completely systematically present for a relation). Alternatively, softer boundaries would suggest that these KNs play a more polysemantic and nuanced role, whereby knowledge is partially distributed across different neurons on different occasions (e.g., the concept of *Paris* and *Madrid* cannot be disentangled at the neuron level, or the relation *capital of* is not always encoded in the same way).

Multilingual Knowledge Neurons Similarly, we ask whether knowledge is language-agnostic; for example, humans do not need to relearn facts when acquiring a new language. Knowledge could be language-dependent in LLMs however: if a fact is present from the English corpus but missing from a Spanish training corpus, an LLM may be able to retrieve that knowledge when prompted in English but not when prompted in Spanish. We employ the KNs framework to investigate the open question of whether a common language-agnostic knowledge representation exists in multilingual models at the level of neurons.

We hypothesize that some KNs may be specific to one language, while others may be sensitive to prompts in multiple languages. We thus analyze the number of languages across which such neurons are shared. We do so by identifying KNs for relations in the `ParaRel` dataset across multiple languages, using the `Multi-ParaRel` dataset, which was specifically created for this multilingual evaluation.

5 Monolingual Experiments: Tracking Concept and Relation Neurons

5.1 Experimental Settings

Models We studied BERT (Devlin et al., 2019), and more precisely `bert-base-uncased` and `bert-large-uncased`, as it has been the reference model for evaluation on TREx since Petroni et al. (2019). Having been trained on Wikipedia from which TREx is derived, their performance is very good (P@1> 0.4). We also studied OPT (Zhang et al., 2022) in its 350 million-parameters version `opt-350m` and 6.7 billion-parameters version `opt-6.7b`, Llama 2 (Touvron et al., 2023) in its 7 billion-parameter version `Llama-2-7b-hf` as well as Gemma 2 (Team et al., 2024) in its 9 billion parameters version `gemma-2-9b`. For all these models we use the HuggingFace implementation. KNs computations were performed on NVIDIA Tesla V100 GPUs for models with less than a billion parameters, and on NVIDIA Tesla A100 GPUs for larger models. The computation took less than an hour per relation.

Template filtering Per model, we excluded prompts with less than 10% top-1 accuracy (that is, accuracy of the most probable continuation). We then excluded relationships with less than 4 prompts left. Since all actual answers were made of a single token, we also limit answers made of a single token. After this filtering,

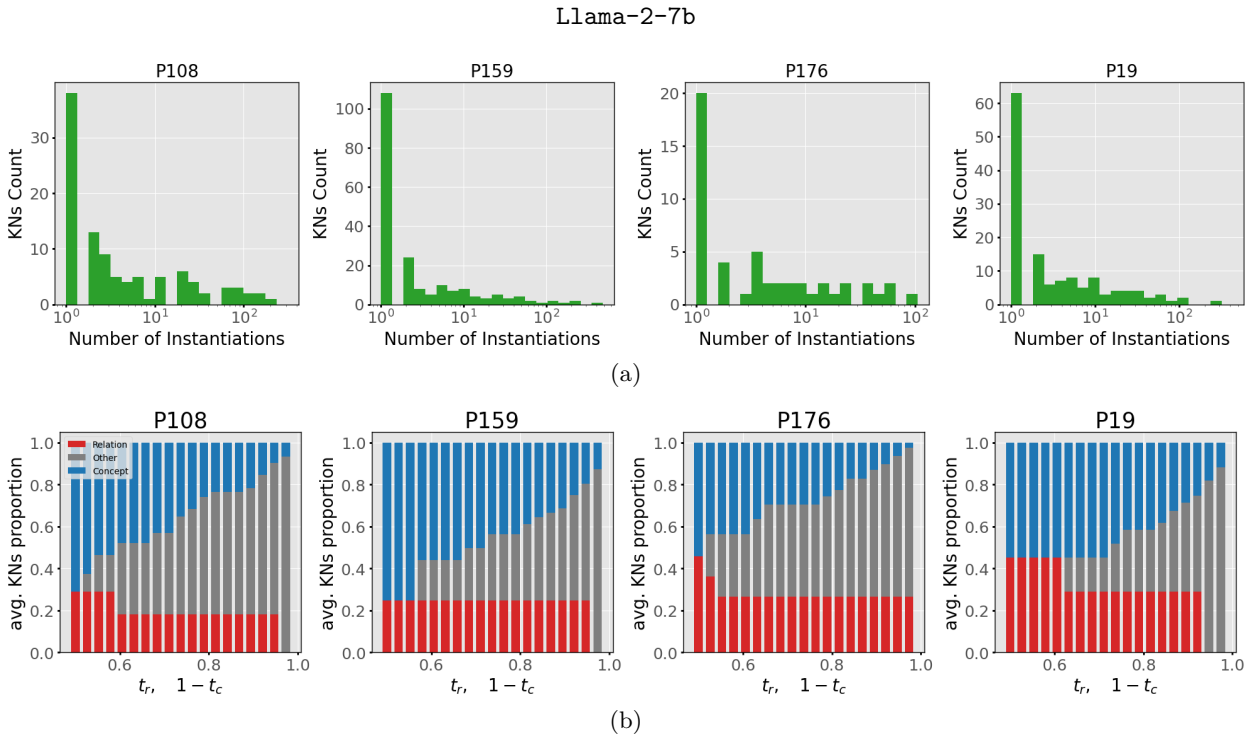


Figure 2: Each panel corresponds to a relation (P108, P159, etc.). (a) Distribution of KNs based on the number of instantiations (i.e. specific triplets, specific facts) within a relation for which a KN was identified. A large number of neurons are identified as KN for a single instantiation, while a roughly similar number of neurons are identified as KN for a continuously increasing number of instantiations within a relation. (b) Average proportion of the KNs from a single instantiation which can be categorized as **Concept**, **Relation** or neither, according to different thresholds (x-axis). The proportion of relational neurons is stable across different thresholds, the proportion of concept neurons decreases with more demanding thresholds.

we obtained on average 15 prompts per relation for BERT and 8 prompts per relation for OPT (starting from 18), confirming the higher accuracy of BERT at the task.

5.2 Tracking a Typology of Knowledge

Before classifying Knowledge Neurons (KNs) according to our typology, we first analyzed the distribution of KNs based on the number of instantiations for which a KN was identified. Figure 2a illustrates the results for four relations using the `Llama-2-7b` model (complete results are provided in Appendix B). A qualitative analysis reveals two key findings: (i) many KNs appear in only one instantiation, indicating that these neurons are task-specific and sensitive to a single concept; and (ii) there is a continuous range of KNs sensitive to between 3 and N instantiations, suggesting a more nuanced role for these neurons that lies between relational and conceptual.

The second observation challenges the simplistic interpretation of assigning neurons exclusively to concepts. At the same time, it also demonstrates the presence of a significant number of neurons sensitive to enough instantiations to hypothesize a more relational role in knowledge retrieval mechanisms.

Thus, we have identified potential candidates for the roles of both **Concept Neurons** and **Relation Neurons**, as well as neurons that fall into an intermediate category. The natural question that arises is: what is the proportion of each neuron type per instantiation, based on thresholds t_r and t_c? This information is not directly inferable from Figure 2a, as neurons appearing consistently across instantiations are less visible

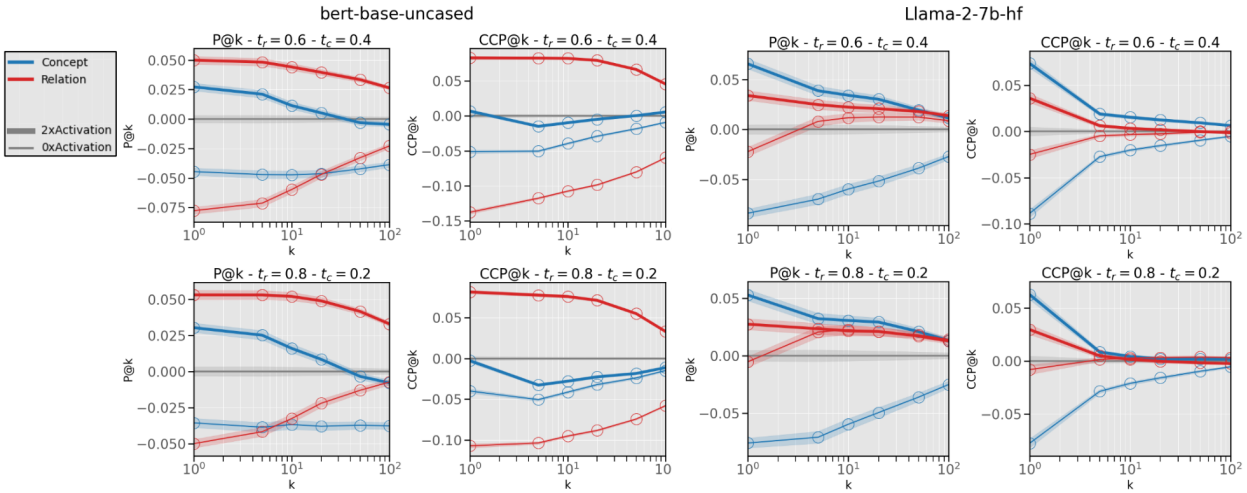


Figure 3: Boosting experiments results for `bert-base-uncased` (left) and `Llama-2-7b` (right) for two couple of thresholds $t_r = 0.6$, $t_c = 0.4$ (top) and $t_r = 0.8$, $t_c = 0.2$ (bottom). The lines corresponds to the ΔP@k (resp. ΔCCP@k) for different k values ranging from 1 to 100. Thick lines represents the doubled activations results and thin lines the nullified activations results. We also plotted the standrad error accross the evaluated instantiations of the relations.

than neurons that appear uniquely in each instance.[2] To address this, we computed the proportion of each neuron type as a function of thresholds at the instantiation level (see Figure 2b). For simplicity, we used symmetrical thresholds, setting $t_r = 1 - t_c$.

As expected, when the thresholds become more restrictive, the number of neurons with well-defined roles decreases, giving way to neurons with less clearly defined functions across all relations. For the `Llama-2-7b` model, we observe that the number of neurons classified as **Relation Neurons** remains more stable compared to those classified as **Concept Neurons**. Furthermore, for a single instantiation, there are few KNs that are exclusive to that instance: when $t_c < 0.1$, the proportion of **Concept Neurons** is less than 0.2.

We also examined the distribution of neuron types across the model's layers but found no significant variation. As observed by Dai et al. (2022), KNs are primarily concentrated in the final layers.

In summary, we have demonstrated the existence of neurons reacting specifically to a single concept within a relation. We have also identified neurons that play a much broader role in such relations, with some reacting to almost all instances of that relation. We attempt to verify this hypothesis through causal experiments in the next section. Finally, some neurons are activated by a subset of the instantiations, carrying a much less transparent type of knowledge. In principle, it could encode subtypes of relations, such as 'capital of a European country', although we find this highly stipulative at the moment. In the next section, we will focus on concept and relation neurons and evaluate their role through causal experiments.

5.3 Boosting Experiments

In this experiment, we investigate the effect of either doubling or nullifying the activation of KNs on model predictions. Dai et al. (2022) conducted similar experiments, focusing on how manual changes to neuron activations influenced output probabilities. In contrast, we employ two more concrete impact metrics: precision at rank k, denoted P@k, which measures the proportion of correct responses in the top k model predictions, and correct category proportion at rank k, denoted CCP@k, which reflects the proportion of responses in the correct category (e.g., *capitals*) within the top k predictions. The original metric of relative probabilities

[2]For example, if each instantiation contains 10 KNs, including 2 perfect conceptual neurons and 8 perfect relational neurons (present in only 1 instantiation and all instantiations, respectively), Figure 2a would display a bar of 200 at the 1 abscissa and a bar of 8 at the 100 abscissa, which would obscure the predominant role of **Relation Neurons**.

change would not show specificity (e.g. unrelated tokens could be even more boosted). For this reason, we report P@k and CCP@k. Effects here ensure that the boost to the correct answer overcomes any boost for other answers. We also include a control experiment in Appendix B to better investigate specificity.

Our goal is to verify whether the behavior of the identified KNs aligns with our proposed typology. Specifically, we hypothesize that (i) there will be a marked increase (or decrease) in precision at rank k=1 when the activations of **Concept Neurons** are doubled (or nullified), with the effect diminishing as k increases. Similarly, we anticipate (ii) that the effect of **Relation Neurons** on P@k will be weaker than that of **Concept Neurons**, as precision is primarily sensitive to the correct response. In contrast, for the CCP@k metric, we expect (iii) that **Relation Neurons** will play a more significant role, as these neurons should be more likely to favor the correct category (e.g., *capitals*), even if it does not boost the correct answer specifically. We assess these effects for a range of thresholds t_c and t_r. Results for the `bert-base-uncased` and `Llama-2-7b` models are shown in Figure 3 (see Appendix B for additional models and thresholds as well).

The figures show the delta in P@k and CCP@k for the predictions with altered (doubling or nulifying) vs unaltered activations. The horizontal line at zero thus represents the baseline model performance. Of the six models evaluated, all six display the expected effect (i) consistently across all thresholds: in short, the top response is more accurate when the activations of concept neurons are increased. However, only two models, Llama-2 and the Gemma-2, exhibit effect (ii). Additionally, four models, belonging to the BERT and OPT families, align with expectation (iii). Overall, `bert-large-uncased` and `gemma-2-9b` adhere to all three expected behaviors across all cases. This happens under restrictive thresholds however ($t_r = 0.9$ and $t_c = 0.1$), and the four other tested models fail to match all of these expectations.

These mixed results show that classifying KNs into distinct and disentangled roles is not perfect, potentially due to noise in our methods or in knowledge attribution methods in the first place. Yet, our experiments do indicate that, for certain models, KNs exhibit specific behaviors and manipulating them leads to predictable effects.

5.4 Discussion

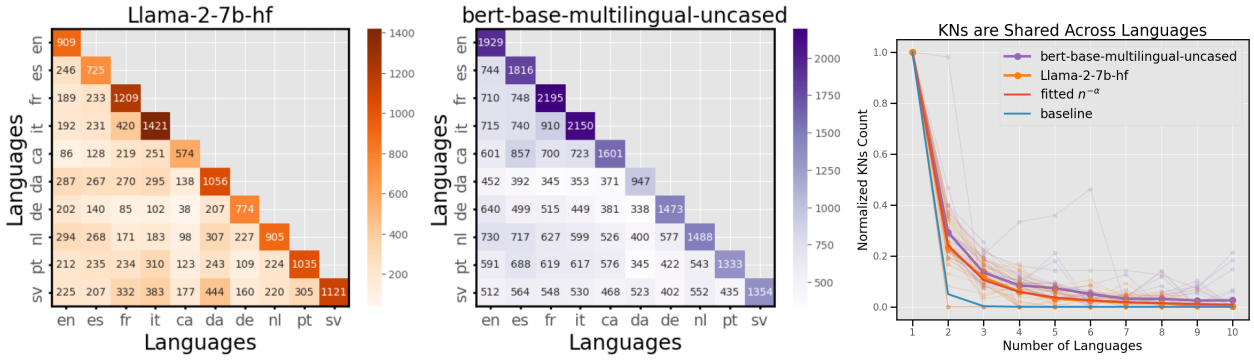
As anticipated, these experiments underscore the complexity of the internal mechanisms within LLMs, making it impractical to map a single, well-defined function to individual neurons. Many of the identified KNs do not adhere to a clearly defined role and cannot be neatly categorized as encoding either concepts or relations, even within a highly controlled environment like `ParaRel`. This is consistent across all the models studied. We believe that the polysemantic nature of neurons prevents such precise delineation, which also helps explain the knowledge editing limitations highlighted in prior research. However, contrary to our initial expectations, certain KNs do appear to serve rather specific functions, and this has been experimentally confirmed for some models in boosting experiments. Nonetheless, we observed significant variation in behavior for the different models, which tends to demonstrate that the observed effects are sometimes fragile. Hence, while the idea that knowledge would be represented entirely in mono-semantic single neurons is unrealistic, the historically associated methods of, e.g., Knowledge Neurons nonetheless detect transparent signal about how knowledge is encoded. KNs are thus a useful tool to pursue the study of knowledge representation in multilingual models too, which we do in the next section.

6 Multilingual Experiments

When we learn a new language, we do not learn all facts about the world again, just new ways to express them. That is, there is a central knowledge base, that we can prompt with several languages. In this section we inquire if knowledge is shared across languages in multilingual models too and, if so, what knowledge.

6.1 Experimental Settings

Models For this experiment we studied `bert-base-multilingual-uncased` (Devlin et al., 2019) and `Llama-2-7b`. We used a NVIDIA Tesla V100 GPU for BERT and NVIDIA Tesla A100 GPU for Llama 2, both for about one hour per relation and per language.



(a) (b) (c)

Figure 4: (a) Number of KNs shared by language pairs for `Llama-2-7b`. About a quarter of neurons are shared between two languages. (b) Same for `bert-base-multilingual-uncased`. (c) Proportion of shared KNs in a relation as a function of the number of languages in the intersection for `Llama-2-7b` and `bert-base-multilingual-uncased`.

Multi-ParaRel We built and release a new dataset `Multi-ParaRel`, a multilingual version of `ParaRel`. More details are given in Appendix A. `Multi-ParaRel` currently includes 10 languages: English, French, Spanish, Catalan, Danish, German, Italian, Dutch, Portuguese and Swedish. We also offer a translation and curation pipeline which makes it possible to add more paraphrases and more languages. It has an average of 17 prompts per relation and per language but this value varies (from 9 for German to 19 for English). Each prompt is compatible with autoregressive models. After filtering for quality as above, we obtain on average 10 prompts per relation and language.

6.2 Knowledge Neurons are Shared Across Languages

Are KNs Bilingual? KNs were calculated separately for each relation and language. A KN is considered shared between two languages if it appears as a KN in both languages for the same relation. We conducted this pairwise analysis across all languages, thereby extending the findings of Chen et al. (2024) to encompass 10 languages.

The results are presented in Figures 4a and 4b. For the `Llama-2-7b` model, over a quarter of the neurons are shared between any two languages, with this proportion increasing to approximately one-third for `bert-base-multilingual-uncased`. This represents a significant degree of neuron sharing, especially when considering that `bert-base-uncased`, for example, has more than $12 \times 3,072 = 36,864$ neurons in the intermediate layers of its FFNs. To quantify this, note that among these $36,864$ neurons, only $1,929$ are identified as KNs across all relations for English, and $2,195$ for French (roughly 5%). If KNs were randomly selected for each language, we would expect around 100 shared neurons between them (5% overlap); however, in reality, 710 neurons are shared. A similar analysis for `Llama-2-7b` gives even more extreme results: by chance, there should be 2 shared neurons, while in practice 189 are found. Moreover, these numbers represent a lower bound, as some relations were excluded from the prompt filtering process for certain language pairs, effectively reducing the shared KN count for those relations to zero. Thus, the data indicates significant overlap of KNs across languages, suggesting a partially shared mechanism for knowledge retrieval across different language pairs.

Are KNs Multilingual? Next, we examine how the number of shared KNs scales with the number of languages in the intersection. Figure 4c shows these results for all relations, along with the average behavior. Across all relations, we observe a consistent pattern: the number of shared neurons decays as a function of the form $(\text{number of languages})^{-\alpha}$, with a fitted $\alpha = 2.04$ for `Llama-2-7b`. In comparison, if neurons were shared at random, the expected behavior would follow $\propto p^{\text{number of languages}}$, where p is the probability of a neuron being a KN (e.g. $p = 0.05$ for BERT). This demonstrates that KNs are more multilingual than

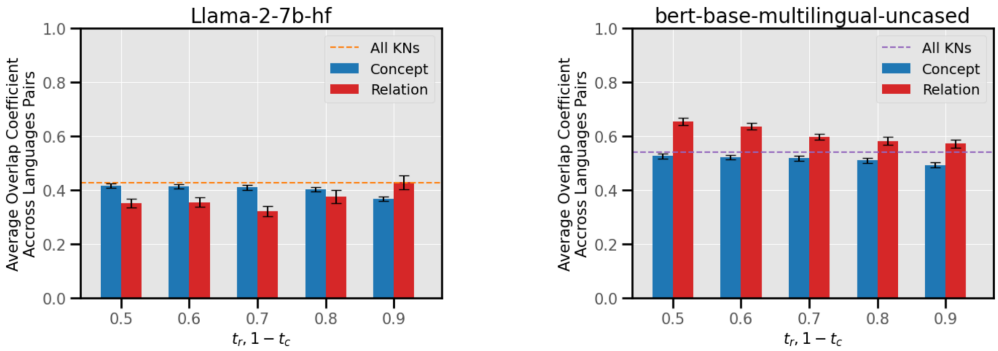


Figure 5: Influence of typology on the average overlap coefficient calculated per language pair of `Llama-2-7b` (left) and `bert-base-multilingual-uncased` (right).

chance, reinforcing the notion of a language-agnostic knowledge retrieval mechanism. Similar to the findings in Section 5, we observe some but few neurons activated for *all* languages.

Are some neurons more Multilingual? **Concept Neurons** and **Relation Neurons** were computed separately for each language and each model. Figure 5 displays the average pairwise overlap coefficient for each neuron type, across various t_r and t_c thresholds, alongside with the pairwise overlap coefficient for all KNs. The results reveal a significant difference in overlap between **Concept Neurons** and **Relation Neurons** at all threshold levels. However, the direction of this difference varies depending on the model and on the threshold. At the most demanding thresholds (those to the right selecting the purest types), we observe that relational neurons appear to be more bilingual. Given the variability at other thresholds (in particular for Llama, which is less performant than BERT in this task), we remain cautious about this conclusion.

6.3 Multilingual Boosting Experiments

While knowledge neurons may be shared across languages, this does not guarantee that they serve the same role in the two languages. A neuron active in both English and French for a given task may perform different overall tasks depending on the language, that is, parallel activation does not equate to shared functionality.

In this section, we conduct a boosting experiment—similar to the one in Section 3—to assess whether neurons shared across languages have a similar causal effect on the model's output. Specifically, for a given language pair, we identify the shared neurons and then either nullify or double their activations when the model is prompted in each language. The goal is to determine whether these neurons are more sensitive to one language more than the other.

Figure 6 presents the results for English and French. As expected, doubling the activation leads to higher P@k and CCP@k scores, while nullifying it results in a decrease. Importantly, the magnitude of these effects is similar across both languages. When activations are nullified, neither language shows a clear advantage. Doubling the activations gives English a slight edge, particularly for lower values of k, but the difference is minor relative to the overall effect. These findings suggest that the shared neurons exert a comparable causal influence in both languages, indicating a similar functional role.

6.4 Knowledge Neurons are Shared Between Natural and Unnatural Languages

We have extended the analysis to non-natural languages, in order to deepen the work of Kervadec et al. (2023) in the specific framework of KNs. More specifically, we calculated 10 seeds of Autoprompt (Shin et al., 2020) for each model and each relation of `ParaRel` and the associated KNs. We then calculated the overlap coefficient between the KNs calculated in this way and those calculated for English at the relationship level. The results are presented in Table 1. This reveals a very large overlap for all models, going up to an almost complete overlap ($\geq 80\%$) for models other than BERT. In the same way that there were important

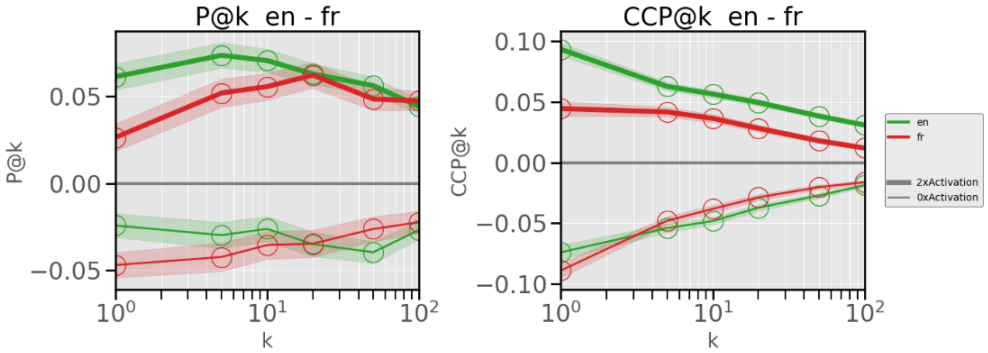


Figure 6: Multilingual Boosting experiments results for `bert-base-multilingual-uncased` for English and French.The lines corresponds to the ΔP@k (resp. ΔCCP@k) for different k values ranging from 1 to 100. Thick lines represents the doubled activations results and thin lines the nullified activations results. We also plotted the standrad error accross the evaluated instantiations of the relations. Here only shared KNs' activations are modified.

overlaps across natural languages, this new result suggests a similar mechanism of knowledge retrieval even between natural and non-natural languages. It is possible however that there exists a confound here because both Autoprompt and KNs are gradient based.

Model	`bert-base`	`bert-large`	`opt-350m`	`opt-6.7b`	`Llama-2-7b`
Avg. Overlap Coeff.	40%	32%	83%	87%	79%

Table 1: Average overlap coefficient of KNs sets computed at the relation level between English and Autoprompt.

7 Discussion and Limitations

Relying on KNs represents a limitation of our work. Their identification relies on an attribution method based on gradient computations, as opposed to alternative approaches such as causal tracing. However, we do not identify any obvious bias introduced by such a choice and we argue that our primary contribution lies in the methodology itself. Importantly, the core of our analysis is agnostic to the specific technique used to identify neurons, and could be applied to other attribution methods just as well. We plan such analysis in future works.

Another possible limitation would be the use of threshold to identify **Concept Neurons** and **Relation Neurons**. We argue that this is not the case. First, thresholds are applied after the identification of KNs, not as part of their initial computation. Second, since the functional roles of individual neurons are unknown a priori, we systematically explored a range of thresholds (see Figure 2a, 2b, 5, 12, 13 & 14). This exhaustive approach is an integral part of the method, allowing us to analyze KN behavior across different sensitivity levels. Importantly, we observed no significant variation in results across thresholds, which retroactively supports the validity of the method.

Another limitation of our methodology is that the identification of neurons shared across languages for the same relations and concepts does not guarantee that these neurons serve the same functional role. We explore this issue in Section 6.3, where we present evidence suggesting that some shared neurons may indeed fulfill similar functions. However, further investigation is required to systematically examine different intersections, whether at the level of relations, concepts, responses, or prompt formats, in order to more precisely determine these roles. Our approach provides a useful framework for refining such analyses and uncovering functionally shared neurons across languages.

Finally, the neurons identified in this study were discovered within a highly controlled setting, specifically using the TREx dataset. As such, the identification of KNs, as well as the proposed categories of **Concept Neurons** and **Relation Neurons**, may not directly translate to real-world scenarios in their current form. We plan to explore ways of adapting this methodology to more naturalistic contexts in future work.

8 Conclusion

We introduced a typology for knowledge and applied it to the knowledge attribution method proposed by Dai et al. (2022) to better classify and understand the behavior of Knowledge Neurons (KNs). Notably, our method remains agnostic to the specific knowledge attribution technique used. Coherently with the initial assumptions in the original work, we found that some of these neurons encode specific concepts, but we also found many which do not and instead seem to exhibit a distributed role, where multiple neurons share responsibility for encoding concepts within the same relation, or maybe encode the whole relation. We hypothesize that this polysemantic nature of neurons contributes to the mixed success observed when using KNs for knowledge editing tasks. Yet again, we were able to identify a subset of more specialized neurons, which we categorized as either conceptual (sensitive to a single concept) or relational (sensitive to relationships between concepts). And in some contexts their manual manipulations show the expected effects on downstream tasks. We extended our analysis to multilingual models and found that a significant number of KNs are shared across languages—both in pairwise comparisons and across all 10 languages tested. This indicates the presence of a shared, language-agnostic knowledge base within multilingual models. To facilitate this research, we created a multilingual dataset of facts and prompts, enriched with paraphrases in 10 languages. Our findings suggest that even a simple method like Knowledge Neurons can provide valuable insights into the benefits of multilingual training.

Looking ahead, we aim to further explore how this shared knowledge can be leveraged to improve the integration of new languages into existing multilingual models. Our results indicate that it may not be necessary to relearn factual knowledge for each language, which could pave the way for more efficient training strategies, particularly for low-resource languages. Instead of focusing on exhaustive coverage of world knowledge, future efforts could prioritize data that highlights the unique syntactic and linguistic features of these languages, thus optimizing resource use and improving model performance. Another possible research direction involves examining how the mechanisms of factual knowledge develop throughout training, which could shed light on the most advantageous stages to introduce or integrate external knowledge.

Acknowledgments

We wish to thank Paul Egré, Olivier Pietquin, Benjamin Spector, Tessa Verhoef, Bernhard Weiss, Yair Lakretz, Nur Lan, Shane Steinert-Threlkeld, Dieuwke Hupkes, Corentin Kervadec, Marco Baroni and members of the EVIL group at UPF, and of the LINGUAE and LINGUAE ML groups at ENS. We also gratefully acknowledge the support from: ANR-10-IDEX-0001-02, FrontCog (ANR-17-EURE-0017), ComCogMean (Projet-ANR-23-CE28-0016), and a gift from Meta. This work was performed using HPC resources from GENCI–IDRIS (Grant AD011013783).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, and et al. Gpt-4 technical report. (arXiv:2303.08774), March 2024. doi: 10.48550/arXiv.2303.08774. URL `http://arxiv.org/abs/2303.08774`. arXiv:2303.08774 [cs].

Templeton Adly, Conerly Tom, Marcus Jonathan, Lindsey Jack, Bricken Trenton, Chen Brian, Pearce Adam, Citro Craig, Ameisen Emmanuel, Jones Andy, Cunningham Hoagy, L Turner Nicholas, McDougall Callum, MacDiarmid Monte, Tamkin Alex, Durmus Esin, Hume Tristan, Mosconi Francesco, Freeman C. Daniel, R. Sumers Theodore, Rees Edward, Batson Joshua, Jermyn Adam, Carter Shan, Olah Chris, and Henighan Tom. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. May 2024. URL `https://transformer-circuits.pub/2024/scaling-monosemanticity/`.

Roee Aharoni, Melvin Johnson, and Orhan Firat. Massively multilingual neural machine translation. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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. 3874–3884, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1388. URL https://aclanthology.org/N19-1388.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. Findings of the 2021 conference on machine translation (WMT21). In Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz (eds.), *Proceedings of the Sixth Conference on Machine Translation*, pp. 1–88, Online, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.1.

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. A review on language models as knowledge bases. (arXiv:2204.06031), April 2022. doi: 10.48550/arXiv.2204.06031. URL http://arxiv.org/abs/2204.06031. 99 citations (Semantic Scholar/arXiv) [2024-04-29] 99 citations (Semantic Scholar/DOI) [2024-04-29] arXiv:2204.06031 [cs].

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. Massively multilingual neural machine translation in the wild: Findings and challenges. (arXiv:1907.05019), July 2019. doi: 10.48550/arXiv.1907.05019. URL http://arxiv.org/abs/1907.05019.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3632–3642, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1399. URL https://aclanthology.org/D18-1399.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. Building machine translation systems for the next thousand languages, 2022. URL https://arxiv.org/abs/2205.03983.

David Bau, Steven Liu, Tongzhou Wang, Jun-Yan Zhu, and Antonio Torralba. Rewriting a deep generative model. (arXiv:2007.15646), July 2020a. doi: 10.48550/arXiv.2007.15646. URL http://arxiv.org/abs/2007.15646. arXiv:2007.15646 [cs].

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 117(48):30071–30078, December 2020b. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1907375117. arXiv:2009.05041 [cs].

Rachel Bawden and François Yvon. Investigating the translation performance of a large multilingual language model: the case of BLOOM. In Mary Nurminen, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz (eds.), *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pp. 157–170, Tampere, Finland, June 2023. European Association for Machine Translation. URL https://aclanthology.org/2023.eamt-1.16.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Journey to the center of the knowledge neurons: Discoveries of language-independent knowledge neurons and degenerate knowledge neurons. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(1616):17817–17825, March 2024. ISSN 2374-3468. doi: 10.1609/aaai.v38i16.29735.

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Knowledge localization: Mission not accomplished? enter query localization!, 2025. URL https://arxiv.org/abs/2405.14117.

Rochelle Choenni, Dan Garrette, and Ekaterina Shutova. How do languages influence each other? studying cross-lingual data sharing during LM fine-tuning. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13244–13257, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.818. URL https://aclanthology.org/2023.emnlp-main.818/.

Lynn Chua, Badih Ghazi, Yangsibo Huang, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Amer Sinha, Chulin Xie, and Chiyuan Zhang. Crosslingual capabilities and knowledge barriers in multilingual large language models, 2025. URL https://arxiv.org/abs/2406.16135.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models. (arXiv:2307.12976), December 2023. doi: 10.48550/arXiv.2307.12976. URL http://arxiv.org/abs/2307.12976. arXiv:2307.12976 [cs].

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 Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, July 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. Emerging cross-lingual structure in pretrained language models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6022–6034, Online, July 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.536. URL https://aclanthology.org/2020.acl-main.536.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *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. doi: 10.18653/v1/2022.acl-long.581. URL https://aclanthology.org/2022.acl-long.581.

Nicola De Cao, Wilker Aziz, and Ivan Titov. Editing factual knowledge in language models. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6491–6506, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.522. URL https://aclanthology.org/2021.emnlp-main.522.

Nicola De Cao, Leon Schmid, Dieuwke Hupkes, and Ivan Titov. Sparse interventions in language models with differentiable masking. In Jasmijn Bastings, Yonatan Belinkov, Yanai Elazar, Dieuwke Hupkes, Naomi Saphra, and Sarah Wiegreffe (eds.), *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 16–27, Abu Dhabi, United Arab Emirates (Hybrid),

December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.blackboxnlp-1.2. URL https://aclanthology.org/2022.blackboxnlp-1.2.

Ameet Deshpande, Partha Talukdar, and Karthik Narasimhan. When is BERT multilingual? isolating crucial ingredients for cross-lingual transfer. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3610–3623, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.264. URL https://aclanthology.org/2022.naacl-main.264.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021a. doi: 10.1162/tacl_a_00410. URL https://aclanthology.org/2021.tacl-1.60.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021b. doi: 10.1162/tacl_a_00410. 224 citations (Semantic Scholar/DOI) [2024-04-29].

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. T-REx: A large scale alignment of natural language with knowledge base triples. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1544.

Xavier Garcia, Aditya Siddhant, Orhan Firat, and Ankur Parikh. Harnessing multilinguality in unsupervised machine translation for rare languages. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1126–1137, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.89. URL https://aclanthology.org/2021.naacl-main.89.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446.

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517, 2023. doi: 10.1162/tacl_a_00615. URL https://aclanthology.org/2023.tacl-1.85/.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. (arXiv:2301.04213), October 2023. doi: 10.48550/arXiv.2301.04213. URL http://arxiv.org/abs/2301.04213. 46 citations

(Semantic Scholar/arXiv) [2024-02-19] 46 citations (Semantic Scholar/DOI) [2024-02-19] arXiv:2301.04213 [cs].

Linyang He, Peili Chen, Ercong Nie, Yuanning Li, and Jonathan R. Brennan. Decoding probing: Revealing internal linguistic structures in neural language models using minimal pairs. (arXiv:2403.17299), March 2024. doi: 10.48550/arXiv.2403.17299. URL http://arxiv.org/abs/2403.17299. arXiv:2403.17299 [cs, q-bio].

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. How good are gpt models at machine translation? a comprehensive evaluation, 2023. URL https://arxiv.org/abs/2302.09210.

Jing Huang, Atticus Geiger, Karel D'Oosterlinck, Zhengxuan Wu, and Christopher Potts. Rigorously assessing natural language explanations of neurons. (arXiv:2309.10312), September 2023. doi: 10.48550/arXiv.2309.10312. URL http://arxiv.org/abs/2309.10312. 5 citations (Semantic Scholar/arXiv) [2024-03-24] 5 citations (Semantic Scholar/DOI) [2024-03-24] arXiv:2309.10312 [cs].

Kuan-Hao Huang, Wasi Ahmad, Nanyun Peng, and Kai-Wei Chang. Improving zero-shot cross-lingual transfer learning via robust training. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1684–1697, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.126. URL https://aclanthology.org/2021.emnlp-main.126.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021. doi: 10.1162/tacl_a_00407. URL https://aclanthology.org/2021.tacl-1.57.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. Cross-lingual ability of multilingual bert: An empirical study. (arXiv:1912.07840), February 2020. doi: 10.48550/arXiv.1912.07840. URL http://arxiv.org/abs/1912.07840. arXiv:1912.07840 [cs].

Nora Kassner, Philipp Dufter, and Hinrich Schütze. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *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. doi: 10.18653/v1/2021.eacl-main.284. URL https://aclanthology.org/2021.eacl-main.284.

Corentin Kervadec, Francesca Franzon, and Marco Baroni. Unnatural language processing: How do language models handle machine-generated prompts? pp. 14377–14392, Singapore, December 2023. doi: 10.18653/v1/2023.findings-emnlp.959. URL https://aclanthology.org/2023.findings-emnlp.959.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6919–6971, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.384. URL https://aclanthology.org/2024.naacl-long.384.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. Investigating multilingual NMT representations at scale. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1565–1575, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1167. URL https://aclanthology.org/D19-1167.

Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. The emergence of number and syntax units in LSTM language models. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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. 11–20, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1002. URL https://aclanthology.org/N19-1002.

Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. (arXiv:1901.07291), January 2019. doi: 10.48550/arXiv.1901.07291. URL http://arxiv.org/abs/1901.07291. arXiv:1901.07291 [cs].

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. February 2018. URL https://openreview.net/forum?id=rkYTTf-AZ.

Danni Liu and Jan Niehues. Learning an artificial language for knowledge-sharing in multilingual translation. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri (eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 188–202, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.12.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, November 2020. ISSN 2307-387X. doi: 10.1162/tacl_a_00343.

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054, December 2020. doi: 10.1073/pnas.1907367117.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. (arXiv:2202.05262), January 2022. URL http://arxiv.org/abs/2202.05262. arXiv:2202.05262 [cs].

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. (arXiv:2210.07229), August 2023. URL http://arxiv.org/abs/2210.07229. 171 citations (Semantic Scholar/arXiv) [2024-03-01] arXiv:2210.07229 [cs].

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale. (arXiv:2110.11309), June 2021. URL http://arxiv.org/abs/2110.11309. 188 citations (Semantic Scholar/arXiv) [2024-04-29] arXiv:2110.11309 [cs].

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. Memory-based model editing at scale. (arXiv:2206.06520), June 2022. doi: 10.48550/arXiv.2206.06520. URL http://arxiv.org/abs/2206.06520. 125 citations (Semantic Scholar/arXiv) [2024-03-04] arXiv:2206.06520 [cs].

Jingcheng Niu, Andrew Liu, Zining Zhu, and Gerald Penn. What does the knowledge neuron thesis have to do with knowledge? October 2023. URL https://openreview.net/forum?id=2HJRwwbV3G.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. Towards making the most of ChatGPT for machine translation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5622–5633, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.373. URL https://aclanthology.org/2023.findings-emnlp.373/.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2463–2473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250. URL https://aclanthology.org/D19-1250.

Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. How context affects language models' factual predictions. (arXiv:2005.04611), May 2020. doi: 10.48550/arXiv.2005.04611. URL http://arxiv.org/abs/2005.04611. 166 citations (Semantic Scholar/arXiv) [2024-04-29] arXiv:2005.04611 [cs].

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1493. URL https://aclanthology.org/P19-1493.

Maxime Poli, Thomas Schatz, Emmanuel Dupoux, and Marvin Lavechin. Modeling the initial state of early phonetic learning in infants. *Language Development Research*, 5(1):1–34, 2024. Published: 2024/04/23.

Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. Learning to generate reviews and discovering sentiment. (arXiv:1704.01444), April 2017. doi: 10.48550/arXiv.1704.01444. URL http://arxiv.org/abs/1704.01444. arXiv:1704.01444 [cs].

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. URL https://api.semanticscholar.org/CorpusID:160025533.

Sara Rajaee and Christof Monz. Analyzing the evaluation of cross-lingual knowledge transfer in multilingual language models. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2895–2914, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-long.177/.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. The curious case of hallucinations in neural machine translation. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1172–1183, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.92. URL https://aclanthology.org/2021.naacl-main.92.

Ryokan Ri and Yoshimasa Tsuruoka. Pretraining with artificial language: Studying transferable knowledge in language models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7302–7315, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.504. URL https://aclanthology.org/2022.acl-long.504.

Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5418–5426, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.437. URL https://aclanthology.org/2020.emnlp-main.437.

David Robinson, Marius Miron, Masato Hagiwara, and Olivier Pietquin. Naturelm-audio: an audio-language foundation model for bioacoustics. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=hJVdwBpWjt.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020. doi: 10.1162/tacl_a_00349. URL https://aclanthology.org/2020.tacl-1.54.

Tara Safavi and Danai Koutra. Relational World Knowledge Representation in Contextual Language Models: A Review. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1053–1067, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.81. URL https://aclanthology.org/2021.emnlp-main.81.

Shibani Santurkar, Dimitris Tsipras, Mahalaxmi Elango, David Bau, Antonio Torralba, and Aleksander Madry. Editing a classifier by rewriting its prediction rules. (arXiv:2112.01008), December 2021. doi: 10.48550/arXiv.2112.01008. URL http://arxiv.org/abs/2112.01008. arXiv:2112.01008 [cs].

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, and et al. Bloom: A 176b-parameter open-access multilingual language model. (arXiv:2211.05100), June 2023. doi: 10.48550/arXiv.2211.05100. URL http://arxiv.org/abs/2211.05100. arXiv:2211.05100 [cs].

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In Katrin Erk and Noah A. Smith (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 86–96, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1009. URL https://aclanthology.org/P16-1009.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4222–4235, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.346. URL https://aclanthology.org/2020.emnlp-main.346.

Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong. BERT is not an interlingua and the bias of tokenization. In Colin Cherry, Greg Durrett, George Foster, Reza Haffari, Shahram Khadivi, Nanyun Peng, Xiang Ren, and Swabha Swayamdipta (eds.), *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pp. 47–55, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-6106. URL https://aclanthology.org/D19-6106.

Anton Sinitsin, Vsevolod Plokhotnyuk, Dmitriy Pyrkin, Sergei Popov, and Artem Babenko. Editable neural networks. (arXiv:2004.00345), July 2020. doi: 10.48550/arXiv.2004.00345. URL http://arxiv.org/abs/2004.00345. 110 citations (Semantic Scholar/arXiv) [2024-04-29] arXiv:2004.00345 [cs, stat].

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. (arXiv:1703.01365), June 2017. doi: 10.48550/arXiv.1703.01365. URL http://arxiv.org/abs/1703.01365. arXiv:1703.01365 [cs].

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. Language-specific neurons: The key to multilingual capabilities in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5701–5715, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.309. URL https://aclanthology.org/2024.acl-long.309.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo,

Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size. (arXiv:2408.00118), August 2024. doi: 10.48550/arXiv.2408.00118. URL http://arxiv.org/abs/2408.00118. arXiv:2408.00118 [cs].

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. (arXiv:2307.09288), July 2023. doi: 10.48550/arXiv.2307.09288. URL http://arxiv.org/abs/2307.09288. arXiv:2307.09288 [cs].

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. Prompting PaLM for translation: Assessing strategies and performance. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15406–15427, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.859. URL https://aclanthology.org/2023.acl-long.859.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang,

and Yue Zhang. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity, 2023. URL https://arxiv.org/abs/2310.07521.

Weixuan Wang, Barry Haddow, Minghao Wu, Wei Peng, and Alexandra Birch. Sharing matters: Analysing neurons across languages and tasks in llms, 2024a. URL https://arxiv.org/abs/2406.09265.

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Nenkov Georgiev, Rocktim Jyoti Das, and Preslav Nakov. Factuality of large language models: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19519–19529, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1088. URL https://aclanthology.org/2024.emnlp-main.1088/.

Xiaokai Wei, Shen Wang, Dejiao Zhang, Parminder Bhatia, and Andrew Arnold. Knowledge enhanced pretrained language models: A compreshensive survey. (arXiv:2110.08455), October 2021. URL http://arxiv.org/abs/2110.08455. 28 citations (Semantic Scholar/arXiv) [2024-04-29] arXiv:2110.08455 [cs].

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41.

Xu Yuemei, Hu Ling, Zhao Jiayi, Qiu Zihan, XU Kexin, Ye Yuqi, and Gu Hanwen. A survey on multilingual large language models: Corpora, alignment, and bias, 2024. URL https://arxiv.org/abs/2404.00929v3.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models. (arXiv:2205.01068), June 2022. doi: 10.48550/arXiv.2205.01068. URL http://arxiv.org/abs/2205.01068. arXiv:2205.01068 [cs].

A New dataset: Multi-ParaRel

Creation Procedure To build the Multi-ParaRel dataset, we used our augmented autoregressive version of ParaRel and mLAMA. The goal is to translate a template such as *The capital of [X] is [Y]*. The problem is that translators are confused by the presence of placeholders *[X]* and *[Y]*, often resulting in translation errors. To overcome the difficulty, we instantiated *[X]* and *[Y]*, translated the whole sentence with these specific instances, and replaced the instantiations back with placeholders. To do so, we used mLAMA, which contains triplets for over 53 languages.

For example, consider the translation from English into French of the template:

The capital of [X] is [Y]

We use the English triplet *<Great Britain, capital of, London>* to obtain the sentence:

The capital of Great Britain is London

This sentence is then translated into French:

La capitale de la Grande Bretagne est Londres

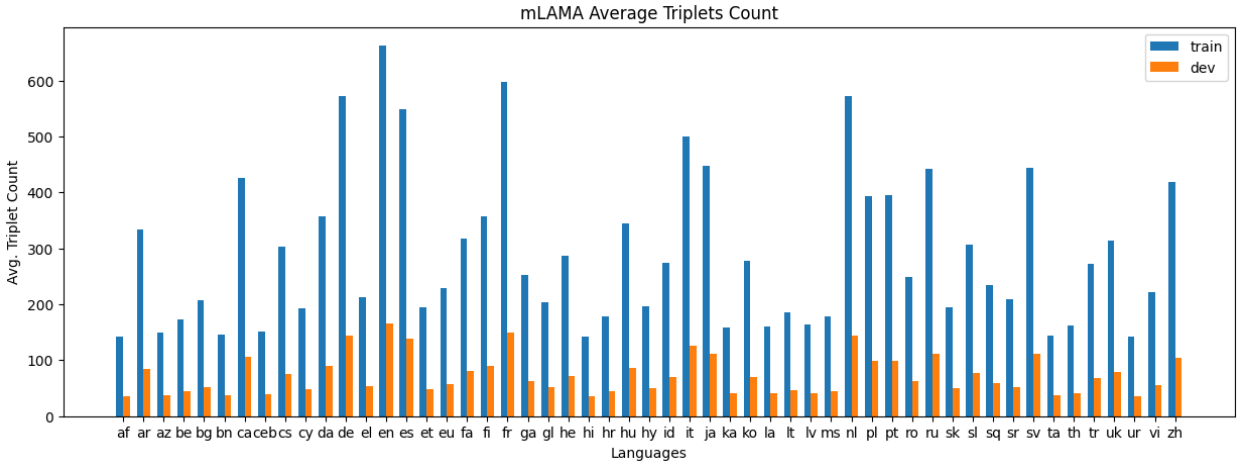


Figure 7: In mLAMA, the number of triplets available varies widely across the different languages.

Then using the French version of the original triplet (*<la Grande Bretagne, capital of, Londres>*), we can find and replace the entity elements of the triplet with placeholders *[X]* and *[Y]*, resulting in the new template:

La capitale de [X] est [Y]

With this overall idea, we can now provide more detail. First of all, such a protocol requires associated triplets in mLAMA from one language to another. However, mLAMA has many more triplets in English than in other languages (see Figure 7), and some triplets are language-specific and therefore cannot be associated with triplets in other languages. We therefore looked into a common English-Target language subset. Then, to avoid translation errors, problems linked to gendered determinants and redundancy (two different templates in English but translated identically in the target language), we used a voting system. Each template was translated 30 times, using 30 triplets. Each translation is assigned a score, which is the number of times the template has been obtained out of the 30 triplets. The template with the highest score is then retained, provided that (i) it is autoregressive, (ii) it has not already been selected and (iii) it is in the top 5 translations.

As a translation model, we used Meta's SeamlessM4T and, more specifically, the Huggingface implementation[3]. We used an NVIDIA Tesla V100 GPU for inference.

Statistics and Exemples Table 3 provides examples of translated templates from different languages and relations. The average number of templates obtained per relationship for each language is:

Quality Analysis To judge the quality of our dataset, we asked a native speaker of French and a native speaker of Spanish to rate the resulting templates in three categories: fluent, weird, ungrammatical. For French 88% are correct, 7% weird and 5% are ungrammatical. For Spanish: 78% of sentences are fluent, 10% weird and 12% are ungrammatical. Although imperfect, `Multi-ParaRel` coupled with a less efficient filtering of prompts gives very good results on mLAMA.

B Full Results

First we provide an overview of all the models behavior with respect to our expectations in Table 4. We also add a control experiment for the BERT family where we conducted the same boosting experiments but sampling KNs randomly within the relation for the Concept Neurons and across relations for Relation Neurons. The goal of such a control is to test the specificity of identified KNs. Results are in Table 5. We see that the effects are destroyed when looking at randomly selected KNs.

[3] `https://huggingface.co/facebook/seamless-m4t-large`

Language	Avg templates
Catalan	19
Danish	15
Dutch	17
English	19
French	14
German	9
Italian	19
Portuguese	19
Spanish	19
Swedish	16

Table 2: Language Values

Relation	English	Spanish	French
P36	*The capital of [X] is [Y]*	*La capital de [X] es [Y]*	*La capitale de [X] est [Y]*
	[X], which has the capital [Y]	*[X], que tiene la capital [Y]*	*[X], dont la capitale est [Y]*
P106	*The occupation of [X] is [Y]*	*La ocupación de [X] es [Y]*	*La profession de [X] est [Y]*
	[X] works as [Y]	*[X] trabaja como [Y]*	*[X] travaille comme [Y]*
P1001	*[X] counts as a legal term in [Y]*	*[X] cuenta como término legal en [Y].*	*[X] est un terme légal en [Y]*
	[X] is a valid legal term in [Y]	*[X] es un término legal válido en [Y].*	*[X] est un terme juridique valide en [Y]*

Table 3: Examples of templates from `Multi-ParaRel`

Model	Expectation (i)	Expectation (ii)	Expectation (iii)
`bert-base-uncased`	Yes	No	Yes
`bert-large-uncased`	Yes	Yes	Yes
`opt-350m`	Yes	No	Yes
`opt-6.7b`	Yes	No	Yes
`Llama-2-7b`	Yes	Yes	No
`gemma-2-9b`	Yes	Yes	No

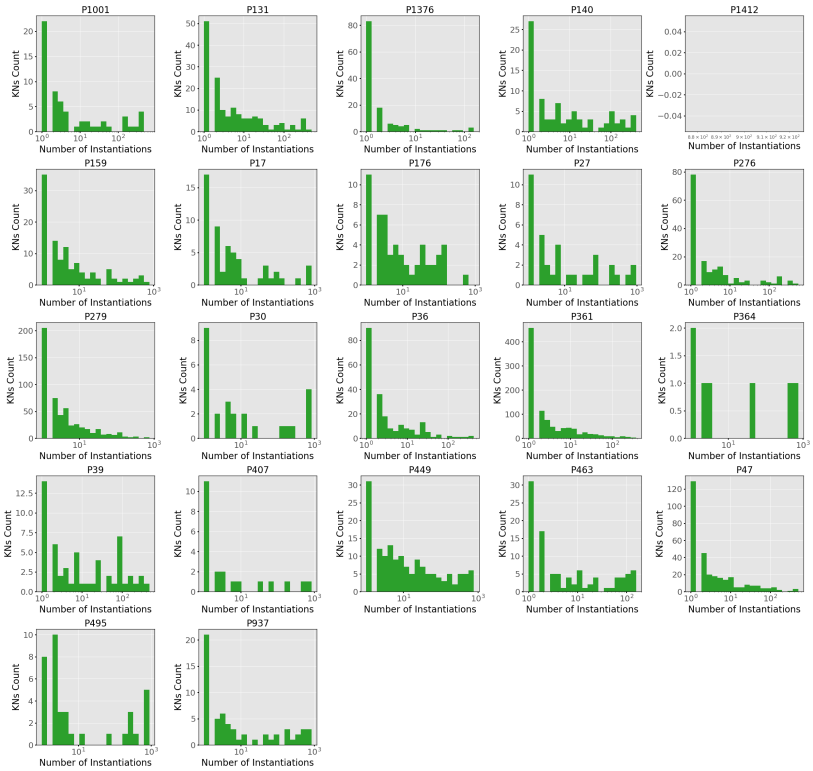
Table 4: Overview of boosting results for all models. Expectations are: (i) there will be a marked increase (or decrease) in precision at rank k=1 when the activations of **Concept Neurons** are doubled (or nullified), with the effect diminishing as k increases, (ii) the effect of **Relation Neurons** on P@k will be weaker than that of **Concept Neurons**, as precision is primarily sensitive to the correct response, (iii) **Relation Neurons** will play a more significant role, as these neurons should be more likely to favor the correct category (e.g., *capitals*), even if it does not boost the correct answer specifically.

Model	Expectation (i)	Expectation (ii)	Expectation (iii)
`bert-base-uncased`	No	No	Yes but effect 10× smaller
`bert-large-uncased`	No	No	No

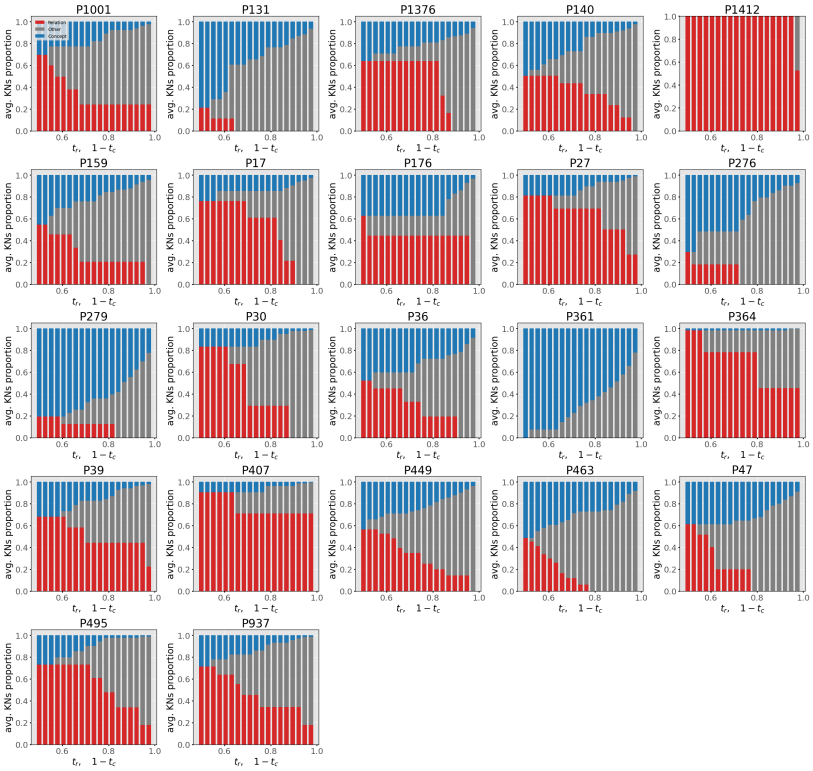
Table 5: Overview of boosting results for the control experiment.

Second, we provide all graphs computed for all models and relations concerning the distinction between concept and relation neurons. This corresponds to the results as presented in Section 5.2, Figure 2, also

showing all relations each time. Second, we provide all graphs corresponding to the boosting experiments (Section 5.3, Figure 3).

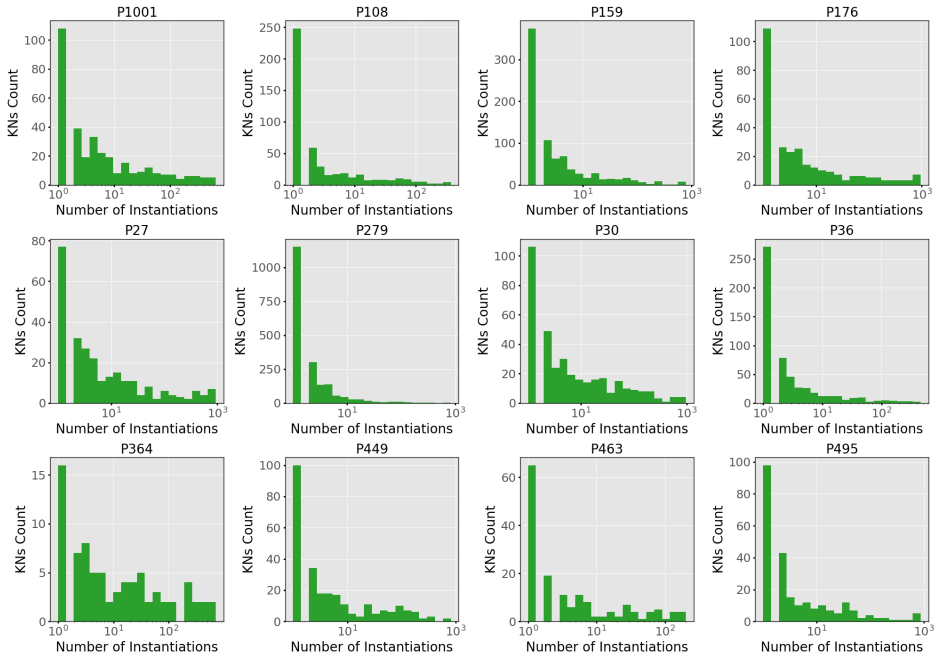


(a)

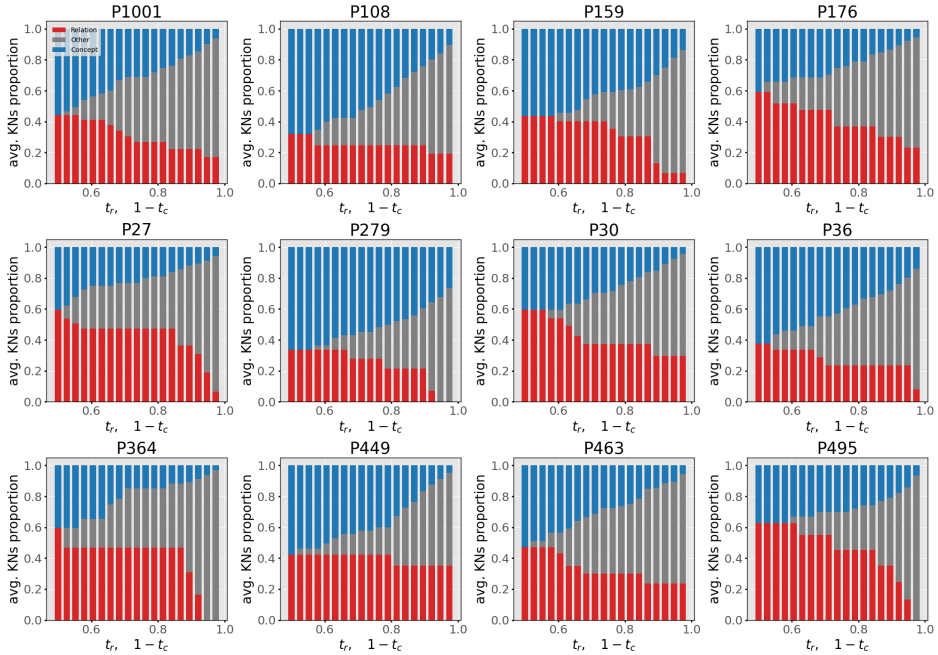


(b)

Figure 8: `bert-base-uncased`

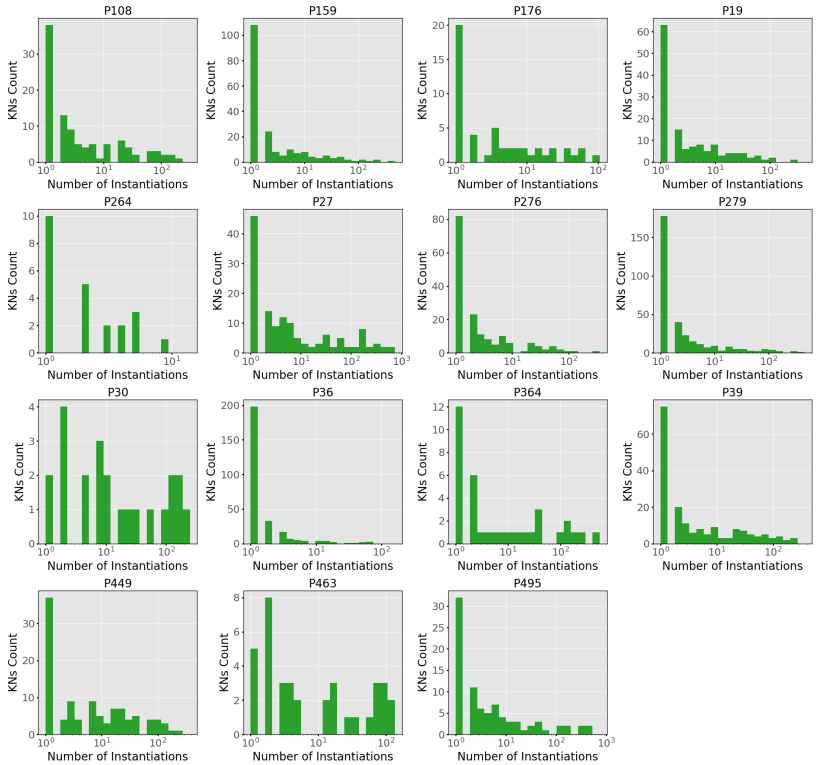


(a)

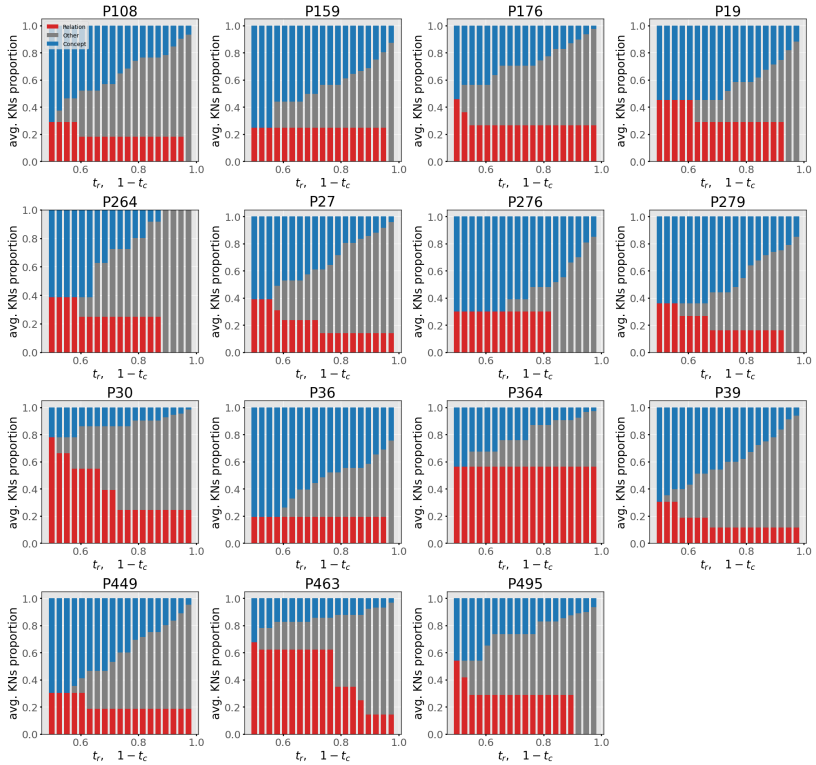


(b)

Figure 9: `opt-6.7b`

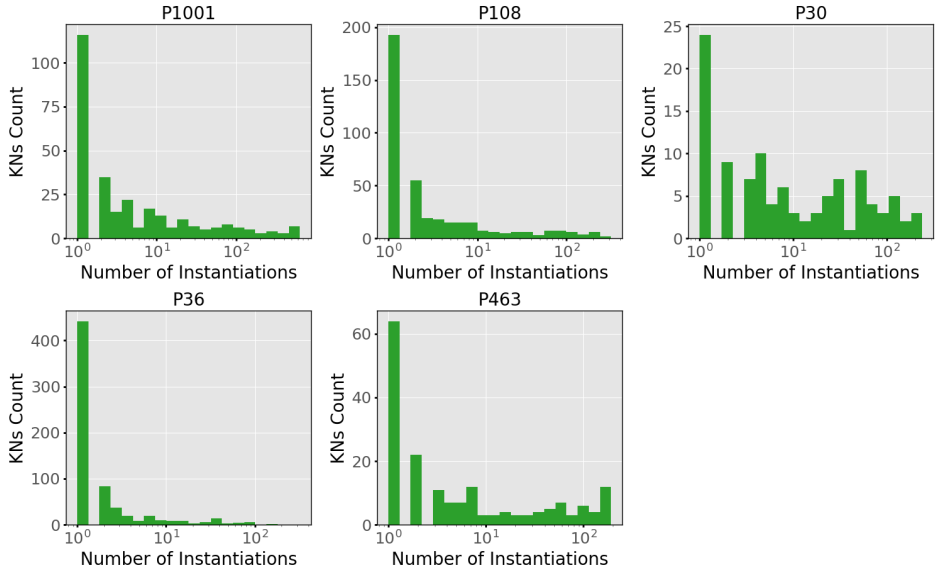


(a)

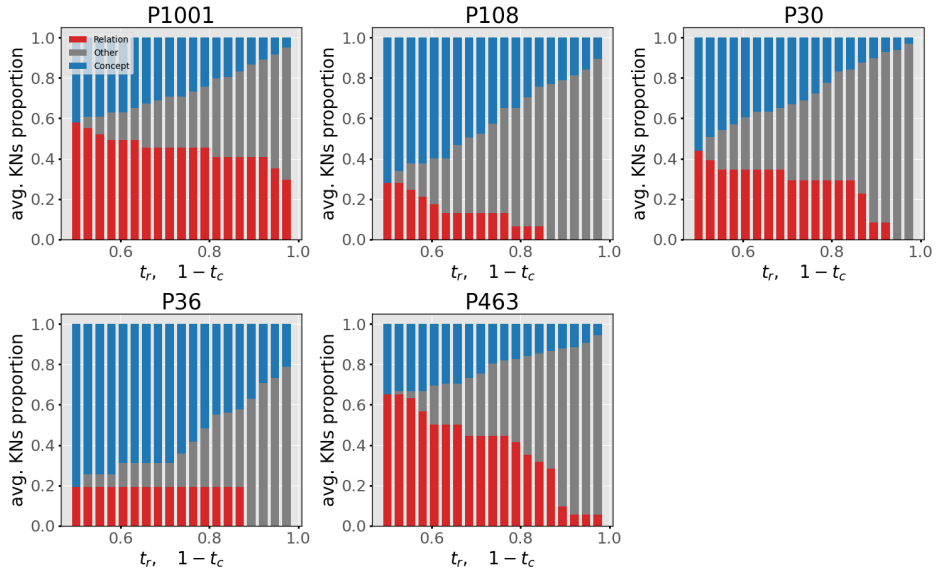


(b)

Figure 10: `Llama-2-7b`



(a)



(b)

Figure 11: `gemma-2-9b`

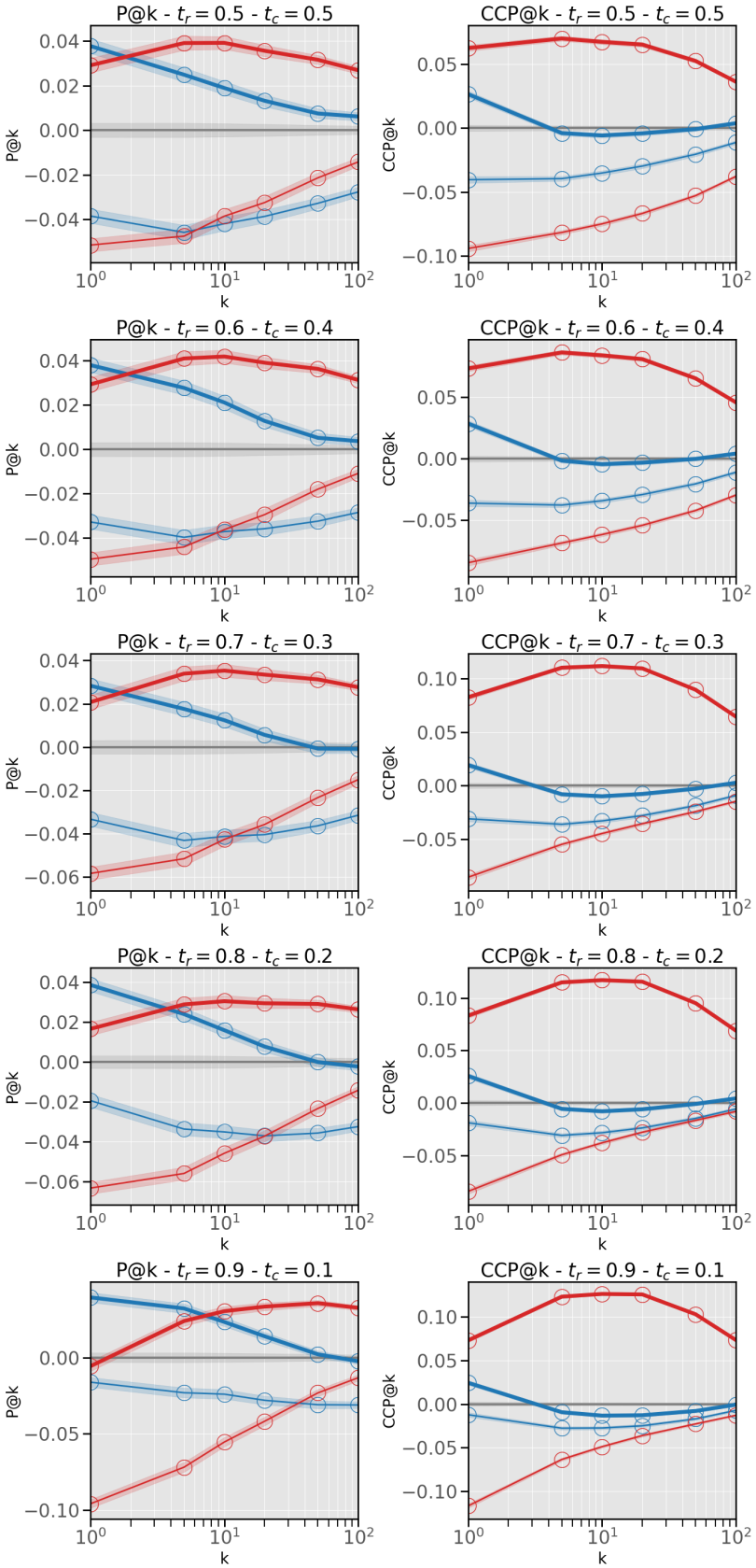


Figure 12: `bert-large-uncased`

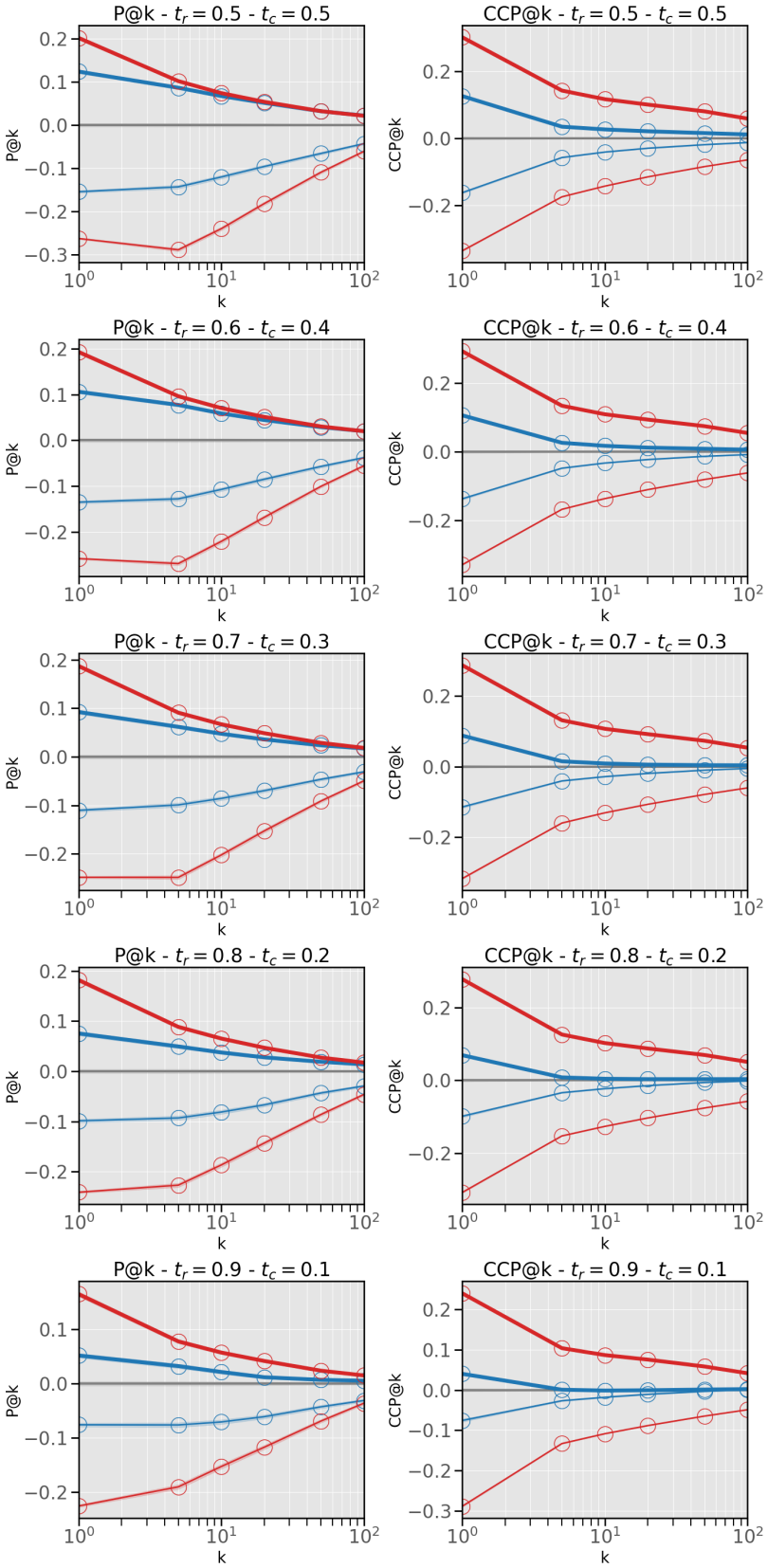


Figure 13: `opt-6.7b`

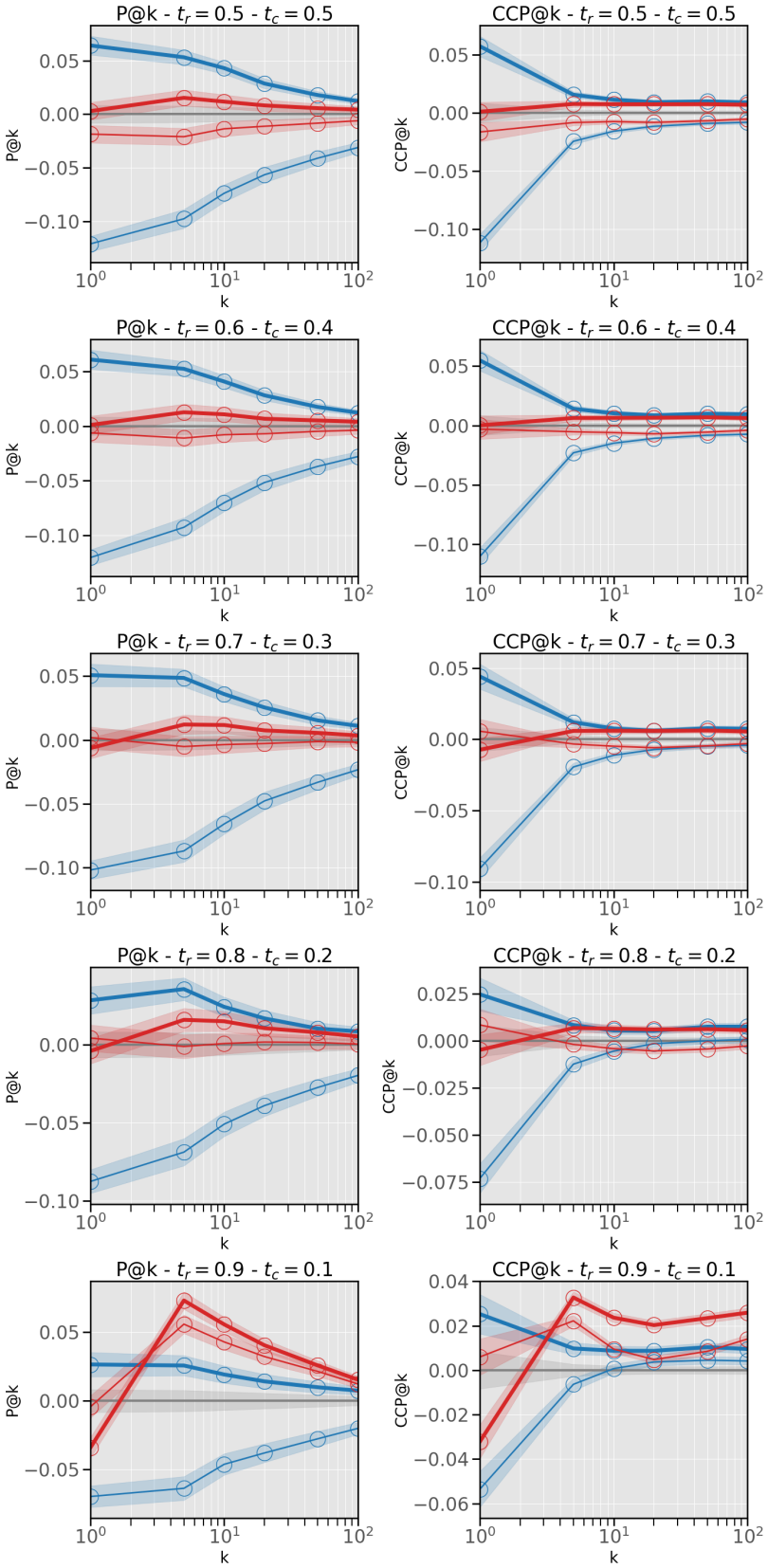


Figure 14: `gemma-2-9b`