Mass-Editing Memory with Attention in Transformers: A cross-lingual exploration of knowledge

Daniel Tamayo[1] **Aitor Gonzalez-Agirre**[1] **Javier Hernando**[1,2] **Marta Villegas**[1]
[1]Barcelona Supercomputing Center
[2]Universitat Politècnica de Catalunya
{daniel.tamayo,aitor.gonzalez,javier.hernando,marta.villegas}@bsc.es

Abstract

Recent research has explored methods for updating and modifying factual knowledge in large language models, often focusing on specific multi-layer perceptron blocks. This study expands on this work by examining the effectiveness of existing knowledge editing methods across languages and delving into the role of attention mechanisms in this process. Drawing from the insights gained, we propose Mass-Editing Memory with Attention in Transformers (MEMAT), a method that achieves significant improvements in all metrics while requiring minimal parameter modifications. MEMAT delivers a remarkable 10% increase in magnitude metrics, benefits languages not included in the training data and also demonstrates a high degree of portability. Our code and data are at https://github.com/dtamayo-nlp/MEMAT.

1 Introduction

Large Language Models (LLMs) based on transformers (Vaswani et al., 2017) are designed to predict the probability of tokens occurring in a sentence rather than comprehending the true semantics that underlie it. As a result, they are susceptible to generating content that lacks a solid grounding in reality and accuracy. Even when two prompts relate to the same factual association $\langle s, r, \cdot \rangle = \langle \mathit{Google}, \mathit{CEO}, \cdot \rangle$; "*The CEO of Google is*" and "*Google's CEO is*", the model lacks an internal constraint that compels it to generate identical answers.

Different investigations have already highlighted the limitations of the models' genuine understanding by analyzing its dependency on the dataset patterns (Gururangan et al., 2018; Jia and Liang, 2017). Furthermore, even when these models seem to *know* the correct answer to a given prompt, they exhibit vulnerability when provided harmful context (Halawi et al., 2023).

In an initial pursuit of exploring the model's true understanding when using knowledge editors, we analyze Mass-Editing Memory in Transformers (MEMIT) (Meng et al., 2023b), a knowledge-editing method asserting its capability to insert up to 10,000 factual associations without heavily inducing catastrophic forgetting. Aligned with previous research (Wang et al., 2023a), our first investigation involves a cross-lingual examination of the limitations associated with MEMIT.

Although the cross-lingual consistency is dependent on the similarity between languages (Qi et al., 2023), our study specifically delves into examining the polyglot capabilities between English and Catalan. In this segment, we construct a translation pipeline to mitigate differences between these languages and proceed to investigate the impact of subject tokenization on knowledge incorporation.

Motivated by the potential of language-independent knowledge neurons (Chen et al., 2023), and the relevance of the attention mechanism in the factual associations domain (Geva et al., 2023), we further our study by exploring a particular part of the attention mechanism: the attention heads. Attention heads have proven to be useful in enhancing the model's reliability under Inference-Time Intervention (ITI) (Li et al., 2023a). The foundational hypothesis behind ITI suggests that attention heads serve as key sources of information for evaluating the truthfulness of models when presented with sentences. Through our experiments, we not only validate the extension of this claim to the domain of factual associations but also observe promising outcomes from a cross-lingual lens. Building on these insights, we propose MEMAT, a method that introduces a novel approach to guide the model towards a better understanding of the edited factual associations.

The proposed method demonstrates improvement across all evaluation metrics from both cross-lingual and monolingual perspectives, showcasing differences exceeding 10% in some cases. Fur-

Findings of the Association for Computational Linguistics ACL 2024, pages 5831–5847
August 11-16, 2024 ©2024 Association for Computational Linguistics

thermore, additional experiments suggest that the modifications introduced by our algorithm enhance the model's understanding of existing knowledge rather than reintroducing it, rendering this approach portable and computationally efficient.

2 Related Work

Retrieval Methods. Rather than directly relying in LLMs for specific queries, open-domain question answering systems has historically been driven by the development of algorithms aligning queries with external database sources (Robertson et al., 2009). Recent advancements in aligning retrieval-based methods with LLMs have demonstrated promise in this domain (Karpukhin et al., 2020; Mao et al., 2020; Borgeaud et al., 2022), with retrieval augmented generation (Lewis et al., 2021) showing capabilities in both multimodal (Chen et al., 2022a,b; Yasunaga et al., 2022) and multilingual (Wang et al., 2023b) contexts. However, while the use of external sources avoids the need for fine-tuning, challenges still persist in precisely identifying the relevant context for a given query (Gao et al., 2023).

Truthfulness. Efforts to enhance the reliability of LLMs without depending on external sources have been a focal point of recent research. Aligning LLMs with human feedback has been explored through Reinforcement Learning from Human Feedback (Stiennon et al., 2020; Ouyang et al., 2022) and Direct Preference Optimization (Rafailov et al., 2023), offering valuable insights for veracity alignment (Chen and Li, 2024). Additionally, approaches contrasting hidden representations of these models have also yielded significant results (Li et al., 2022; Chuang et al., 2023) in this direction.

Factual Knowledge Editors. This research builds upon MEMIT, a method adept at efficiently introducing knowledge by modifying the internal weights of decoder-only architectures, surpassing the effectiveness of earlier meta-learning techniques like MEND (Mitchell et al., 2022a) and constrained fine-tuning (Zhu et al., 2020). Nevertheless, less intrusive alternatives, which selectively modify specific hidden states of the model during inference according to the provided prompt, have also demonstrated remarkable efficacy in knowledge editing. Notable examples include REMEDI (Hernandez et al., 2023), GRACE (Hartvigsen et al., 2022), and SERAC (Mitchell et al., 2022b).

Multilingual Domain. The emergence of knowledge editors and multilingual models raises questions about whether the information is being inserted from a cross-lingual perspective. Current findings suggest that these methods are not entirely language-independent (Schott et al., 2023; Wang et al., 2023a), with approaches based on prompting and retrieval yielding stronger results (Zheng et al., 2023; Wang et al., 2023b).

3 Preliminaries

3.1 Background

Since in our experimental setup English and Catalan were chosen as the languages for conducting experiments, we opted for the utilization of Ǎguila-7B, a decoder-only model consisting of 6.85 billion parameters based on Falcon-7B (Penedo et al., 2023). The internal process performed by this architecture to process text is similar to other decoder-only architectures. It first convert an input to a sequence of N tokens $t_1, t_2, ..., t_N$ by using Byte-level Byte-Pair Encoding (Wang et al., 2020). Then, it process each token by assigning a vector x_i^0 using an embedding matrix $E \in \mathbb{R}^{|\mathcal{V}|} \times d$, where $\mathcal{V}$ denotes the set of vocabulary tokens and d denotes the size of each vector. Following this, the input embeddings undergo a series of L transformer layers, each comprising a Multi-Query Self-Attention (MQSA) sublayer (Shazeer, 2019) and a parallel Multi-Layer Perceptron (MLP) sublayer.

Following the notation proposed in Elhage et al. (2021); Geva et al. (2023), we avoid representing bias terms, layer normalization (Ba et al., 2016), and Rotary Position Embeddings (Su et al., 2023) for simplicity and denote the transformation as:

$$x_i^\ell = x_i^{\ell-1} + a_i^\ell + m_i^\ell, \tag{1}$$

where a_i^ℓ and m_i^ℓ are the outputs from the ℓ-th MQSA and MLP sublayers. In the attention term, for each layer, we assign different projection matrices $W_Q^{\ell,h}, W_K^\ell, W_V^\ell \in \mathbb{R}^{d \times \frac{d}{H}}$ and $W_O^{\ell,h} \in \mathbb{R}^{\frac{d}{H} \times d}$ for $h \in [1, H], \ell \in [1, L]$. Then, given the hidden states of the sentence at layer ℓ denoted as $X^\ell \in \mathbb{R}^{N \times d}$, we define:

$$A^{\ell,h} = \mathcal{S}\left(\frac{(X^{\ell-1}W_Q^{\ell,h})(X^{\ell-1}W_K^\ell)^T}{\sqrt{d/H}} + M^{\ell,h}\right) \tag{2}$$

$$a^\ell = \sum_{h=1}^{H} A^{\ell,h}(X^{\ell-1}W_V^\ell)W_O^{\ell,h}, \tag{3}$$

where $\mathcal{S}$ is a row-wise softmax normalization, and $M^{\ell,h}$ is a mask for $A^{\ell,h}$ that only uses the attention mechanism to modify the token t_r using the previous tokens $t_{\leq r}$ ($M^{\ell,h}_{rc} = -\infty \; \forall c > r$ and zero otherwise).

In the MLP term, we use the matrices $W_{in} \in \mathbb{R}^{d \times d_{ff}}$, $W_{out} \in \mathbb{R}^{d_{ff} \times d}$ and an activation function γ to define:

$$\begin{aligned} k_i^\ell &= \gamma(x_i^{\ell-1} W_{in}) \\ m_i^\ell &= k_i^\ell W_{out}. \end{aligned} \tag{4}$$

3.2 Proposed Framework

While the architecture of large language models is extensively documented, grasping the precise mechanisms that empower them to extract factual information is still matter of research. Notably, studies have revealed the impact of adjusting MLP layers in the generation of factual associations (Geva et al., 2021; Dai et al., 2022a; Chen et al., 2023). This comprehension has paved the way for the development of frameworks such as ROME (Meng et al., 2023a), PMET (Li et al., 2023b), and MEMIT. In the case of PMET and MEMIT, a subset of MLP layers are changed by the introduction of a correction matrix ($\widehat{W}_{out,\ell} = W_{out,\ell} + \Delta_\ell$) such that:

$$\begin{aligned} \widehat{W}_{out,\ell} = \underset{\widetilde{W}_{out,\ell}}{\arg\min}(\sum_{j=1}^{n} ||k_j^\ell \widetilde{W}_{out,\ell} - \widetilde{m}_j^\ell||^2 + \\ \sum_{j=n+1}^{n+u} ||k_j^\ell \widetilde{W}_{out,\ell} - \widetilde{m}_j^\ell||^2). \end{aligned} \tag{5}$$

where n represents the number of factual associations already encoded in the pre-trained model, u represents the number of new factual associations being introduced, each k_i^ℓ is taken from the final position of the subject entities of each factual triplet, and the representation of $\widetilde{m}_j^\ell$ is the one that should be capable of making the model predict the correct factual entity. Refer to Meng et al. (2023b) for more details.

Despite the notable performance of these proposals, recent studies highlight the critical role of attention mechanisms in accurate response generation (Dai et al., 2022b; Yuksekgonul et al., 2023). Specifically, its relevance in factual associations when the attribute extraction is performed (Geva et al., 2023). These findings have already prompted some intervention of attention layers for knowledge editing (Li et al., 2023b; Sakarvadia et al., 2023), a line of research that this study aims to further extend by tailoring attention with the ITI framework.

The core principle of ITI involves a simple method for calculating a subset of head corrections ($\omega^{\ell,h}$) that, when integrated into the language model:

$$\tilde{a}^\ell = \sum_{h=1}^{H} (A^{\ell,h}(X^{\ell-1} W_V^\ell) + \omega^{\ell,h}) W_O^{\ell,h}, \tag{6}$$

result in a significant enhancement in the veracity of the model's responses.

Nevertheless, as knowledge is infused through MEMIT, the notion of truth becomes nuanced. The model can express new information in specific contexts, yet upon closer examination of its reasoning capabilities, a decline in performance is observed (Cohen et al., 2023). The distinctive approach introduced in MEMAT explores how incorporating modifications based on head corrections can optimize the method's understanding of the knowledge introduced while avoiding catastrophic forgetting. Refer to Section 6 for further details.

4 Dataset and Evaluation

The dataset employed in this document is a reduced version of the CounterFact dataset (Meng et al., 2023a,b). For each sample, the relevant prompts utilized in our experiments are:

1. *Efficacy Prompts (EP).* Two distinct objects associated with the same $\langle s, r, \cdot \rangle$ pair, one corresponding to the true fact o^c and the other representing a false fact o^*.

2. *Paraphrase Prompts (PP).* Two prompts that have the same meaning of the $\langle s, r, \cdot \rangle$ pair, but are paraphrased and receive an addition of noise at the beginning. In evaluations, these prompts can also be referred to as indicators of the model's *generalization* capability.

3. *Neighborhood Prompts (NP).* Ten different prompts which contain different subjects ($s_j \neq s$) with the same relation $\langle s_j, r, \cdot \rangle$ that would be true with the object o^c. In evaluations, these prompts are referenced as indicators of the model's ability to *specify* the insertion of knowledge.

As discussed in Schott et al. (2023), a purification of the original dataset is necessary to eliminate sentences with awkward phrasing, consistent inaccuracies, and errors. After performing the refinement proposed, the resulting English dataset consists of 11,550 factual associations. However, since our goal is also to investigate the cross-lingual capabilities of MEMIT and there is no available translated dataset meeting our criteria, we develop a translation pipeline.

In contrast to the methodologies employing Wikidata and Google knowledge graphs for translation verification (Kassner et al., 2021), and utilizing `gpt-3.5-turbo` and `gpt-4` for translations (Wang et al., 2023a), our implementation follows a distinct procedure. Our pipeline translates English sentences to Catalan using `projecte-aina/mt-aina-en-ca`, and maintains the dataset structure with `simalign` (Sabet et al., 2021), an aligner method based on contextual embeddings.

Given the significance of preserving sentence order, cases in which the targets o^c or o^* are translated before the end of the sentences are flagged as errors and discarded if necessary. Additionally, we address a challenge associated with gender differences between English and Catalan. For instance, if the English sentence "*The CTO of OpenAI is*" is translated as "*El CTO d'OpenAI és*", it may introduce bias toward male responses even when the correct answer is o^c = *Mira Murati*. To mitigate this, we create two Catalan samples for each English sample.

Using our pipeline and human supervision, we manage to obtain a reduced version of the CounterFact dataset in English and Catalan containing 11,229 samples. An example of the samples in Catalan can be observed in Appendix A.

To evaluate the performance of the different knowledge editors used in our experiments, this study inherits the evaluation metrics from Meng et al. (2023a,b). For each of the three different prompts contained in the dataset (EP, PP, NP), success and magnitude metrics are defined:

- Success Metrics:

$$\begin{aligned} \text{ES} &:= \mathbb{E}\left[\mathbb{P}[o^*|p] > \mathbb{P}[o^c|p]|p \in EP\right] \\ \text{PS} &:= \mathbb{E}\left[\mathbb{E}_{p\in PP}\left[\mathbb{P}[o^*|p] > \mathbb{P}[o^c|s]\right]\right] \\ \text{NS} &:= \mathbb{E}\left[\mathbb{E}_{p\in NP}\left[\mathbb{P}[o^c|p] > \mathbb{P}[o^*|s]\right]\right]. \end{aligned} \tag{7}$$

- Magnitude Metrics:

$$\begin{aligned} \text{EM} &:= \mathbb{E}\left[\mathbb{P}[o^*|p] - \mathbb{P}[o^c|p]|p \in EP\right] \\ \text{PM} &:= \mathbb{E}\left[\mathbb{E}_{p\in PP}\left[\mathbb{P}[o^*|p] - \mathbb{P}[o^c|s]\right]\right] \\ \text{NM} &:= \mathbb{E}\left[\mathbb{E}_{p\in NP}\left[\mathbb{P}[o^c|p] - \mathbb{P}[o^*|s]\right]\right]. \end{aligned} \tag{8}$$

In prior research (Meng et al., 2023a,b; Li et al., 2023b; Wang et al., 2023a), the assessment of knowledge editors heavily relied on success metrics. Nevertheless, it is crucial to note that a favorable success metric coupled with a low-magnitude metric could suggest uncertainty in the model's confidence regarding the retrieved knowledge. In the following sections, we emphasize the significance of the magnitude metric in uncovering patterns that may not be readily apparent in traditional success metrics.

5 Experiments

Before the introduction of MEMAT, the main aspects that motivate the use of our method are explained in this section. Firstly, Section 5.1 outlines the scope of our analysis and examines the limitations of using only English and Catalan. Our findings suggest a correlation between positive cross-lingual outcomes in MEMIT and higher token similarity between subject tokens, indicating the dependency of our cross-lingual analysis on languages that share subject tokens. Subsequently, in Section 5.2, we evaluate the extent of cross-lingual information in the hidden representations of words by studying attention heads.

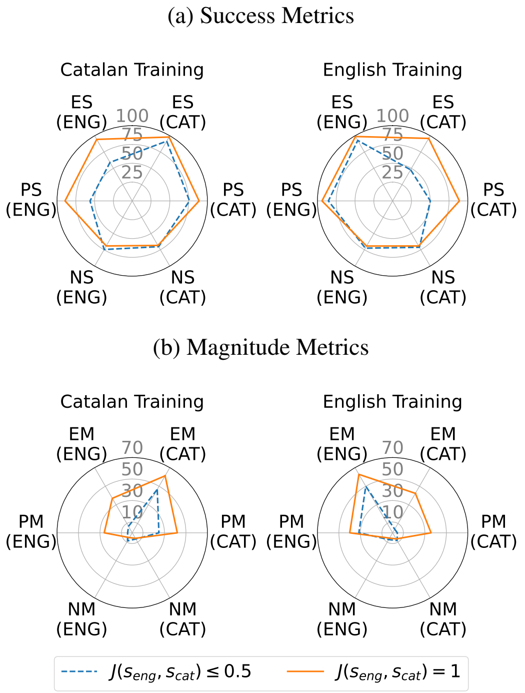
As Ǎguila-7B was not initially assessed using MEMIT, further details on the hyperparameter optimization of the methods studied can be found in Appendix B.

5.1 Cross-Linguality

Considering the substantial resemblance between the English and Catalan alphabets, we investigate the impact of this similarity on the cross-lingual hypotheses asserted in this paper. Utilizing the Jaccard index, expressed as:

$$J(x_{eng}, x_{cat}) = \frac{|x_{eng} \cap x_{cat}|}{|x_{eng} \cup x_{cat}|}, \tag{9}$$

we assess the performance disparity when incorporating factual triplets with distinct subject tokenizations in English and Catalan ($J(s_{eng}, s_{cat}) \leq 0.5$), as opposed to those without such differences ($J(s_{eng}, s_{cat}) = 1$). It is pertinent to note that factual associations typically pertain to entities such as institution names, individuals, and series, which tend to maintain consistent tokenization across languages that share the same alphabet. More details of the exact similarity between both datasets can be found in Appendix C.


(a) Success Metrics

(b) Magnitude Metrics

Figure 1: Results of Efficacy, Generalization and Specificity when applying MEMIT separately in two different languages and evaluating the effects of training in both. Each depicted line show a restriction in the tokenization of the subjects.

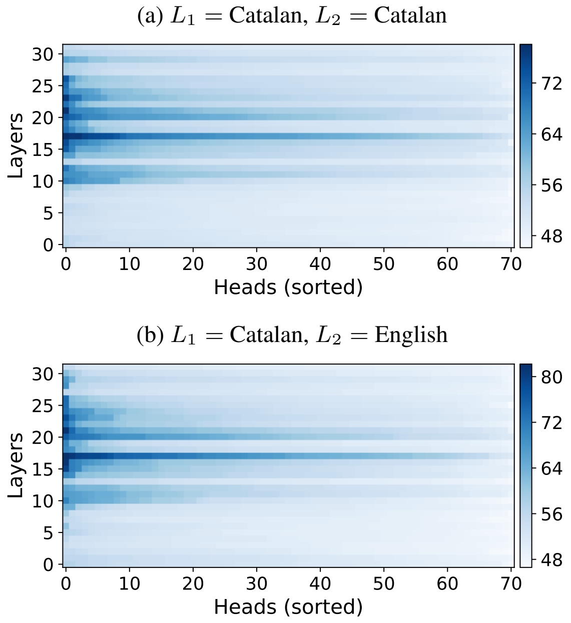
In Figure 1, the outcomes of cross-lingual operations are illustrated for the insertion of 1,000 samples using MEMIT. Two discernible trends emerge from the results:

- Given the dependency of MEMIT in the subject representation, alterations to the subject result in a more pronounced decline in performance from a cross-lingual perspective. This phenomenon may explain the less cross-lingual outcomes observed in Chinese (Wang et al., 2023a).
- When analyzing cases with the same tokenization $(J(s_{eng}, s_{cat}) = 1)$ in both languages, the decline in cross-lingual magnitude metrics is more noticeable than the decrease observed in success metrics. Instances with $(J(s_{eng}, s_{cat}) \leq 0.5)$ experience a significant decrease in performance across both metrics.

5.2 Locating Knowledge with Heads

Under the cross-lingual context outlined in the previous section, we analyze the extent to which the framework of ITI can be useful in the factual knowledge domain. Considering that we have the same set of factual associations in two languages, we denote a language pair as (L_1, L_2) and design the location of knowledge using a specific part of the attention mechanism as follows:

1. We train the model using MEMIT with triplets on language L_1.
2. We identify all heads associated to the last token of M the triplets $\langle s, r, o^c \rangle_i$ and $\langle s, r, o^* \rangle_i$ using L_2, assigning truthful labels $y = 0$ and $y = 1$ respectively[1]. The constructed dataset has the structure: $\{(head^{\ell,h}_{-1,:}, y)_i\}_{i=1}^{M}$, where each head of each layer can be denoted as:
$$head^{\ell,h} = A^{\ell,h}(X^{\ell-1}W_V^{\ell}). \qquad (10)$$
3. We train sigmoid classifiers for each head position (totaling $L \times H$ positions) on subset of triplets, denoted as the training set. The objective is to predict the assigned label by just using a single attention head representation. Subsequently, we utilize the remaining triplets as a validation set to assess the performance of this approach. More implementation details can be found in Appendix D.


(a) L_1 = Catalan, L_2 = Catalan

(b) L_1 = Catalan, L_2 = English

Figure 2: Accuracy on the validation set for all heads in all layers in Ǎguila-7B considering two combinations of L_1 and L_2. The performance peaks include 78.1% and 82.2%. The number of samples introduced using MEMIT is 1,000.

[1] Note that the concept of *truth* in this case is diffuse since the model has already been trained on L_1 and the new true target should be o^*.

If the attention heads could not provide information about whether the sentences are truthful or not, the expected performance should be around a 50% chance of predicting the correct label. However, as highlighted in Li et al. (2023a), empirical observations in a comparable context revealed that certain attention heads achieved 83.3% performance on the validation set in discerning truthful sentences. In this section, our contribution involves not only expanding the application of this framework to MEMIT, but also demonstrating a high degree of language independence. Regardless of the choice of languages L_1 and L_2, some attention heads consistently achieve accurate classification performances near 80%, as shown in Figure 2.

Note that we are just exploring the cross-lingual implications from Catalan to English, but similar patterns can be observed in Appendix E for the converse relation.

6 MEMAT Method

In light of the proven attention heads' relevance, we reinforce the rationale behind equation 6 and present MEMAT as a method that expands upon MEMIT. The overall procedure is depicted in Figure 3, with detailed descriptions for each point as follows:

(a) Firstly, we modify the model with knowledge associated to a set of factual triplets using MEMIT in language L_1, which only edit some MLP layers.

(b) Then, using language L_2, we locate the heads that yield the top K performances using the procedure explained in Section 5.2. Formally, let us consider that, for each classifier learned using the training set, we obtain the predictions $\phi_i^{\ell,h} = H(< head_{-1,\cdot}^{\ell,h}, \theta^{\ell,h} >)$, where H denotes the Heaviside step function and the parameters $\theta^{\ell,h}$ have been trained in the training set. The top head positions can be denoted as those which belong to the set:

$$\Psi^K := \{(\ell,h)|\, \underset{(\ell,h)}{\text{args}\overset{K}{\max}}(\{\phi_i^{\ell,h} \wedge y_i\}_{i=1}^{M\times\beta})\}, \tag{11}$$

where β is the fraction of the validation set.

(c) Finally, under the language L_2, we introduce head corrections $\omega^{\ell,h}$ in each of the K head positions, Ψ^K, and minimize the loss function:

$$\begin{aligned}\mathcal{J}_i^{attn} &= \frac{\lambda_\omega}{K}\sum_{(\ell,h)\in\Psi^K}\left(\frac{||\omega^{\ell,h}||}{||head_{-1,\cdot}^{\ell,h}||}\right)^2 \\ &- \frac{1}{R}\sum_{j=1}^{R}\log \mathbb{P}_{\tilde{G}}[o_i^*|z_j + p(s_i,r_i)] \\ &+ D_{KL}\left(\mathbb{P}_{\tilde{G}}[x|p']||\mathbb{P}_G[x|p']\right),\end{aligned} \tag{12}$$

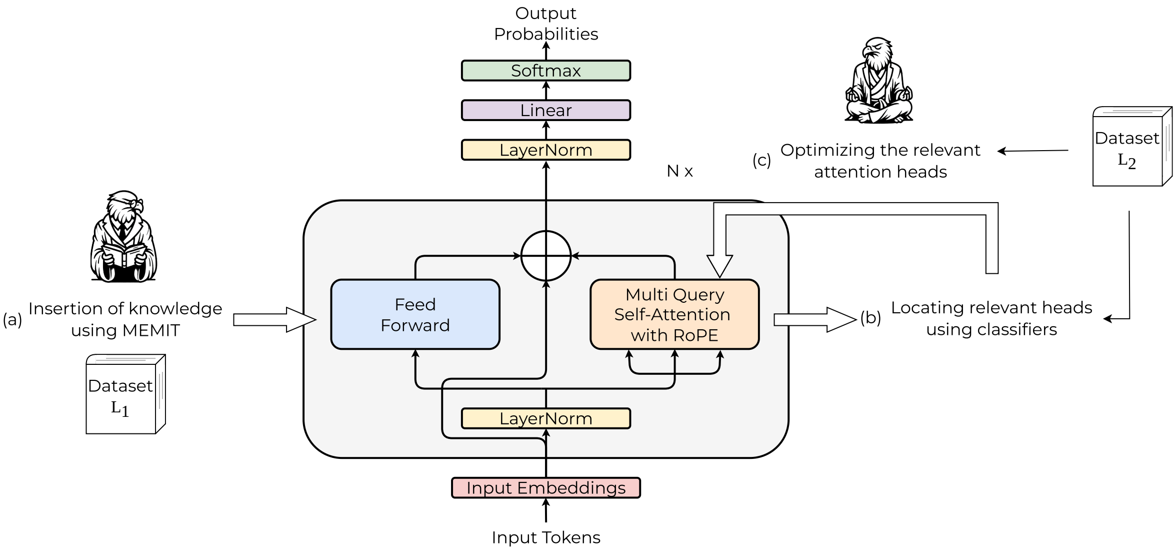

Output Probabilities
Softmax
Linear
LayerNorm
N x
(c) Optimizing the relevant attention heads
Dataset L2
(a) Insertion of knowledge using MEMIT
Dataset L1
Feed Forward
Multi Query Self-Attention with RoPE
(b) Locating relevant heads using classifiers
LayerNorm
Input Embeddings
Input Tokens

Figure 3: Illustration depicting the key steps of MEMAT in Ǎguila-7B. The dataset languages, denoted as L_1 and L_2, are not restricted to differ or remain equal, but in this diagram we consider both datasets to store the same triplets. The Eagle images were generated using GPT-4.

Method (Training Language(s))	English ES	Catalan ES	English PS	Catalan PS	English NS	Catalan NS
Ǎguila-7B Baseline	26.5 (1.2)	25.5 (1.2)	30.4 (1.2)	31.6 (1.2)	**73.9 (0.9)**	72.2 (0.8)
PMET (CAT)	80.4 (2.5)	95.6 (1.3)	72.5 (2.5)	81.3 (2.3)	**73.9 (2.1)**	**74.3 (1.8)**
MEMIT (CAT)	88.8 (0.7)	97.4 (0.4)	84.0 (0.8)	88.3 (0.7)	71.8 (0.8)	70.5 (0.7)
MEMAT-16 (CC)	90.3 (1.1)	**97.8 (0.5)**	86.2 (1.1)	89.6 (1.0)	72.7 (1.2)	71.8 (1.0)
MEMAT-16 (CE)	**90.7 (1.0)**	97.6 (0.6)	**87.0 (1.1)**	89.8 (1.0)	73.4 (1.1)	72.4 (1.0)
MEMAT-16 (CC*)	89.8 (1.1)	97.6 (0.6)	85.7 (1.1)	**90.1 (1.0)**	73.7 (1.1)	72.7 (1.0)
MEMAT-16 (CE*)	89.5 (1.1)	97.3 (0.6)	85.7 (1.1)	89.1 (1.0)	73.1 (1.2)	71.8 (1.0)
	EM	EM	PM	PM	NM	NM
Ǎguila-7B Baseline	-6.7 (0.4)	-7.4 (0.5)	-5.5 (0.4)	-6.2 (0.5)	7.6 (0.3)	8.3 (0.4)
PMET (CAT)	25.3 (2.1)	62.9 (2.0)	23.2 (2.2)	31.9 (2.2)	7.6 (0.7)	8.6 (0.8)
MEMIT (CAT)	31.9 (0.8)	67.0 (0.8)	22.2 (0.6)	40.4 (0.8)	6.6 (0.3)	7.3 (0.3)
MEMAT-16 (CC)	39.0 (1.3)	72.8 (1.2)	27.8 (1.1)	47.9 (1.4)	8.3 (0.5)	9.1 (0.6)
MEMAT-16 (CE)	**43.8 (1.3)**	73.3 (1.2)	**32.2 (1.1)**	50.3 (1.4)	9.7 (0.5)	9.7 (0.6)
MEMAT-16 (CC*)	42.2 (1.4)	**74.8 (1.2)**	31.6 (1.2)	**51.1 (1.4)**	**9.9 (0.5)**	**10.6 (0.6)**
MEMAT-16 (CE*)	38.5 (1.3)	70.6 (1.2)	28.1 (1.1)	46.3 (1.4)	8.7 (0.5)	9.2 (0.5)

Table 1: Results of English and Catalan Efficacy, Generalization and Specificity prompts over the success and magnitude metrics in both languages. Each row represents the experiments performed for the different knowledge editing methods when inserting 1,000 factual associations. The notation assigned to MEMAT-16 is (L_1-L_2), where the cases (L_1-L_2*) indicate the use of attention heads that were trained in a different set of factual triplets and which have been recycled in a new insertion of factual associations. The 95% confidence intervals are in parenthesis.

where $\widetilde{G}$ represents the modified decoder model obtained by inserting $\omega^{\ell,h}$ in the decoder G that results from point (a), ($\widetilde{G} = G([head^{\ell,h}_{-1,\cdot}] += [\omega^{\ell,h}])$). The term D_{KL} is the KL divergence that minimize the effect of the modification in prompts p' that contain the subject s_i, but contain the relationship "*is a*". Finally, R is the number of random prompts z_j that are inserted at the beginning of the sentence to make the optimization more robust under different contexts.

Once the head corrections are optimized, we apply equation 6.

Note that the loss is associated to a single factual triplet. However, differently from MEMIT, we optimize the attention heads corrections for all the samples at the same time by using different batch sizes in Adam (Kingma and Ba, 2017). We also use gradient accumulation to keep the method cheap from a computational perspective.

Given that there is not a clear choice on the number of heads that should be optimized, we make a hyperparameter search on the number of attention heads and find that for most of the metrics, the optimal number of heads is around $K = 16$. For additional information, please refer to Appendix F.

To evaluate the relevance of the information encoded in the attention heads corrections, we also conduct another experiment with our method. We apply MEMAT in a particular set of factual triplets using the pair (L_1, L_2) and save the head corrections and its positions. Then, we insert different factual triplets in the original model using MEMIT in L_1 and add the head corrections that were previously obtained. The results of this experiment, as well as the results of PMET, MEMIT and MEMAT are shown in Table 1 for some combinations of L_1 and L_2, the remaining combinations are left in Appendix G. Additionally, an exploration of experiments that introduce both languages at the same is provided in Appendix H, showing similar results.

Our research indicates that while PMET shows superior performance in neighborhood metrics, this comes at the expense of reduced efficacy and generalization in knowledge introduction, with MEMAT demonstrating promising results in this area. Across all analyzed metrics, MEMAT consistently outperforms MEMIT. Particularly noteworthy are the significant improvements in paraphrase magnitudes, exceeding 10% over the baseline. These findings, combined with positive neighborhood metrics, suggest that optimizing attention heads can improve the comprehension of implicit

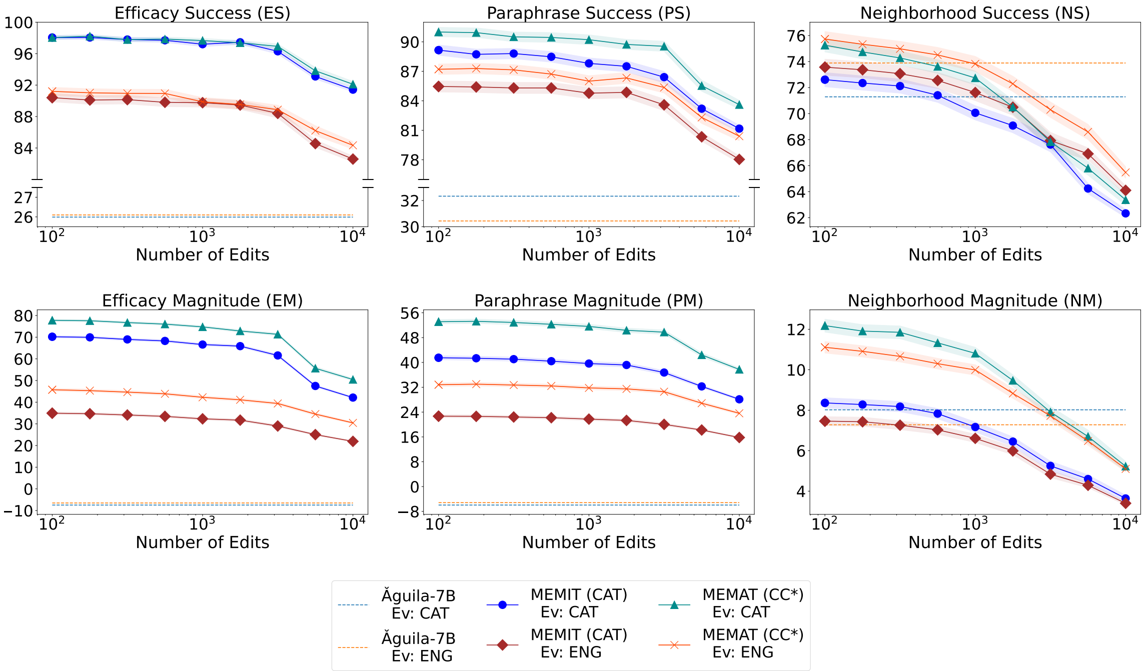


Figure 4: MEMIT and MEMAT scaling curves plot showing the performance of English and Catalan against number of edits (log-scale) when only using Catalan training data. The error correspond to a 68% confidence interval.

knowledge in LLMs. This hypothesis gains further support from our experiments with recycled attention heads on other sets of factual triplets (L_1-L_2* cases), which occasionally outperform the original MEMAT approach.

Moreover, our cross-lingual analysis provides evidence that this enhanced understanding occurs at a certain cross-lingual level, showing positive results in multilingual metrics even when only using monolingual training.

In point (c), the method suggested in ITI was not employed. The decision to deviate from this method stems from observed performance declines within the specified domain. Further details on this matter can be found in Appendix I.

6.1 Scaling Curves

Considering the notable improvement in performance metrics, we opt to conduct a comprehensive comparison of the training evolution between MEMIT and MEMAT in Figure 4. This investigation involves varying the number of inserted samples, with a specific focus on experiments exceeding 100 samples. The sample distribution follows the formula $n_i = \exp(\ln(10,000) * \frac{i}{16})$.

Recognizing the impracticality of training head corrections with only 100 samples, we opt to optimize the head corrections using a subset of 1,000 factual triplets for the combination $L_1 = L_2 = Catalan$. The samples used to train these heads are separated from our dataset. Then, we insert the head corrections into the decoders G_{n_i} that result from inserting different factual triplets with MEMIT in Ăguila-7B. Specifically, MEMAT (CC*) refers to training MEMIT on n_i factual triplets, excluding the initial 1,000, while incorporating the previously obtained attention head corrections.

Although MEMAT's performance still degrades with the introduction of more factual associations, it consistently outperforms MEMIT across all evaluation metrics. This experiment provides additional evidence that head corrections are highly portable and that MEMAT enhances the understanding of previously unseen languages.

7 Reproducibility

The conducted experiments have been executed on workstations equipped with AMD Radeon Instinct MI50 GPUs, with 32 GB of memory each. HuggingFace Transformers (Wolf et al., 2020) facilitates the loading of language models, while PyTorch (Paszke et al., 2019) is employed to implement model editing algorithms on the GPUs. Additionally, the training of sigmoid classifiers is carried out using the Scikit-learn library (Pedregosa et al., 2018) on CPUs.

In this specific setup, introducing 1,000 samples through MEMIT takes 3 GPU hours, contrasting with the 25 GPU minutes required for training 16 attention heads corrections.

8 Conclusions

In this study, we examine the cross-lingual implications of knowledge within the domain of knowledge editors, identifying two significant patterns. Firstly, the proposed methods heavily rely on subject tokenization. Secondly, our experiments show evidence that attention heads encode information in a certain language-independent manner.

Expanding our investigation, we introduce MEMAT, a method that, following the application of MEMIT, fortifies the language model's knowledge through subtle parameter adjustments. We substantiate how the approach introduced is portable and, regardless of the language used during training, enhances the performance of other languages.

9 Future Work

Our work emphasizes the limitations of training LLMs with monolingual data. As a future direction, we are interested in further investigating language adaptation techniques to enable these models to perform tasks in a more language-agnostic manner.

Additionally, we consider necessary to explore the role that each architecture component play in alternative domains. Recognizing the incomplete understanding of transformer-based models, we assert that prioritizing explainable AI could be essential to gain the insights necessary to enhance current state-of-the-art methods. We hope that our study can contribute to inspiring further exploration in this domain.

10 Limitations

All hypotheses put forth in this study stem from experiments conducted in English and Catalan. It is essential to recognize that due to the similarity between their alphabet and the phenomena explored in Section 5.1, the generalization of our findings to other linguistic contexts may be limited. Further experimentation involving diverse languages is imperative to establish the cross-lingual implications of the identified patterns.

Moreover, despite considerable efforts within the natural language processing community, many challenges related to the reliability of language models still persist. The limitations in knowledge editions, as indicated by the inability to modify related knowledge (Cohen et al., 2023) and the lack of bidirectionality (Ma et al., 2023), suggest that exclusively focusing on specific parameters may not offer a solution to the issues of knowledge editing (Pinter and Elhadad, 2023).

Acknowledgements

This work has been promoted and financed by the Government of Catalonia through the Aina project, by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project ILENIA with reference 2022/TL22/00215337, 2022/TL22/00215336, 2022/TL22/00215335, 2022/TL22/00215334, as well as by the Spanish Ministerio de Ciencia e Innovación through the AdaVoice project (PID2019-107579RB-I00).

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Weixin Chen and Bo Li. 2024. Grath: Gradual self-truthifying for large language models. *arXiv preprint arXiv:2401.12292*.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W Cohen. 2022a. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. *arXiv preprint arXiv:2210.02928*.

Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. 2022b. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*.

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2023. Journey to the center of the knowledge neurons: Discoveries of language-independent knowledge neurons and degenerate knowledge neurons.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2023. Evaluating the ripple effects of knowledge editing in language models. *arXiv preprint arXiv:2307.12976*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022a. Knowledge neurons in pretrained transformers.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. 2022b. Why can gpt learn in-context? language models secretly perform gradient descent as meta optimizers. *arXiv preprint arXiv:2212.10559*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data.

Danny Halawi, Jean-Stanislas Denain, and Jacob Steinhardt. 2023. Overthinking the truth: Understanding how language models process false demonstrations.

Thomas Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. 2022. Aging with grace: Lifelong model editing with discrete key-value adaptors. *arXiv preprint arXiv:2211.11031*.

Evan Hernandez, Belinda Z. Li, and Jacob Andreas. 2023. Inspecting and editing knowledge representations in language models.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual lama: Investigating knowledge in multilingual pretrained language models. *arXiv preprint arXiv:2102.00894*.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023a. Inference-time intervention: Eliciting truthful answers from a language model.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2022. Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097*.

Xiaopeng Li, Shasha Li, Shezheng Song, Jing Yang, Jun Ma, and Jie Yu. 2023b. Pmet: Precise model editing in a transformer.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods.

Jun-Yu Ma, Jia-Chen Gu, Zhen-Hua Ling, Quan Liu, and Cong Liu. 2023. Untying the reversal curse via bidirectional language model editing. *arXiv preprint arXiv:2310.10322*.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2020. Generation-augmented retrieval for open-domain question answering. *arXiv preprint arXiv:2009.08553*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023a. Locating and editing factual associations in gpt.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2023b. Mass-editing memory in a transformer.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022a. Fast model editing at scale.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. 2022b. Memory-based model editing at scale.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback, 2022. *URL https://arxiv.org/abs/2203.02155*, 13.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2018. Scikit-learn: Machine learning in python.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only.

Yuval Pinter and Michael Elhadad. 2023. Emptying the ocean with a spoon: Should we edit models? *arXiv preprint arXiv:2310.11958*.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. *arXiv preprint arXiv:2310.10378*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2021. Simalign: High quality word alignments without parallel training data using static and contextualized embeddings.

Mansi Sakarvadia, Aswathy Ajith, Arham Khan, Daniel Grzenda, Nathaniel Hudson, André Bauer, Kyle Chard, and Ian Foster. 2023. Memory injections: Correcting multi-hop reasoning failures during inference in transformer-based language models. *arXiv preprint arXiv:2309.05605*.

Tim Schott, Daniel Furman, and Shreshta Bhat. 2023. Polyglot or not? measuring multilingual encyclopedic knowledge retrieval from foundation language models.

Noam Shazeer. 2019. Fast transformer decoding: One write-head is all you need.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2023. Roformer: Enhanced transformer with rotary position embedding.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2020. Neural machine translation with byte-level subwords. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9154–9160.

Jiaan Wang, Yunlong Liang, Zengkui Sun, Yuxuan Cao, and Jiarong Xu. 2023a. Cross-lingual knowledge editing in large language models.

Weixuan Wang, Barry Haddow, and Alexandra Birch. 2023b. Retrieval-augmented multilingual knowledge editing. *arXiv preprint arXiv:2312.13040*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2022. Retrieval-augmented multimodal language modeling. *arXiv preprint arXiv:2211.12561*.

Mert Yuksekgonul, Varun Chandrasekaran, Erik Jones, Suriya Gunasekar, Ranjita Naik, Hamid Palangi, Ece Kamar, and Besmira Nushi. 2023. Attention satisfies: A constraint-satisfaction lens on factual errors of language models. *arXiv preprint arXiv:2309.15098*.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. Can we edit factual knowledge by in-context learning?

Chen Zhu, Ankit Singh Rawat, Manzil Zaheer, Srinadh Bhojanapalli, Daliang Li, Felix Yu, and Sanjiv Kumar. 2020. Modifying memories in transformer models. *arXiv preprint arXiv:2012.00363*.

A Translated CounterFact Sample

```
{
        "case_id": 2,
        "pararel_idx": 13704,
        "relation_id": "P1303",
        "eval_target_new": {
            "requested_rewrite": {
                "prompt": "{}, el",
                "target_new": {
                    "str": "piano",
                    "id": "Q5994"
                },
                "subject": "Toko Yasuda"
            },
            "paraphrase_prompts": [
               "Inicialment i són zero i és fals.
                ↪  Toko Yasuda, tocant al",
                "La densitat de població era .
                ↪  Toko Yasuda toca el"
            ],
            "neighborhood_prompts": [
                "Paul McCartney toca el",
                "John Lennon, tocant el",
                "Elvis Presley, el",
                "Douglas Adams, tocant el",
                "John Lennon toca el",
                "Jimi Hendrix, tocant el",
                "Ringo Starr, tocant el",
                "Leonard Cohen toca el",
                "Bruce Springsteen, tocant el",
                "John Lennon toca el"
            ],
        },
        "eval_target_true": {
            "requested_rewrite": {
                "prompt": "{}, la",
                "target_true": {
                    "str": "guitarra",
                    "id": "Q6607"
                },
                "subject": "Toko Yasuda"
            },
            "paraphrase_prompts": [
               "Inicialment i són zero i és fals.
                ↪  Toko Yasuda, tocant a la",
                "La densitat de població era .
                ↪  Toko Yasuda toca la"
            ],
            "neighborhood_prompts": [
                "Paul McCartney toca la",
                "John Lennon, tocant la",
                "Elvis Presley, la",
                "Douglas Adams, tocant la",
                "John Lennon toca la",
                "Jimi Hendrix, tocant la",
                "Ringo Starr, tocant la",
                "Leonard Cohen toca la",
                "Bruce Springsteen, tocant la",
                "John Lennon toca la"
            ],
        }
    }
```

B Optimization of the Learning Rate

When applying a hyperparameter search in MEMIT and PMET, we find a stronger effect when changing the learning rate, which we consider optimal at $lr = 0.2$ and $lr = 10^4$ respectively. In Figures 5 and 6, an evolution of the success metric associated to the different prompts is shown given different values of the learning rate. In our search we found MEMIT with superior performance, which denotes some evidence of the dependence of these methods on the model architecture.

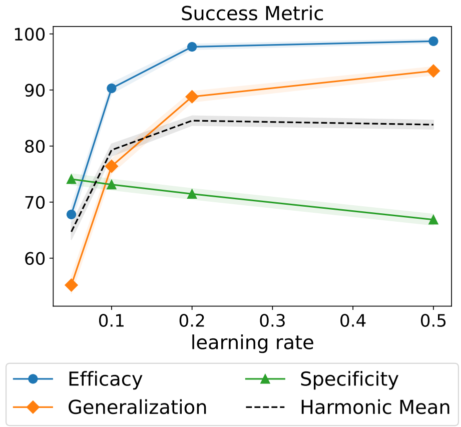


Figure 5: Results over training using MEMIT in Catalan and evaluating in Catalan with different values of the learning rate. Each solid line represents a different type of prompt used, and the dashed line represents the harmonic mean between them. The displayed areas correspond to a 68% confidence interval.

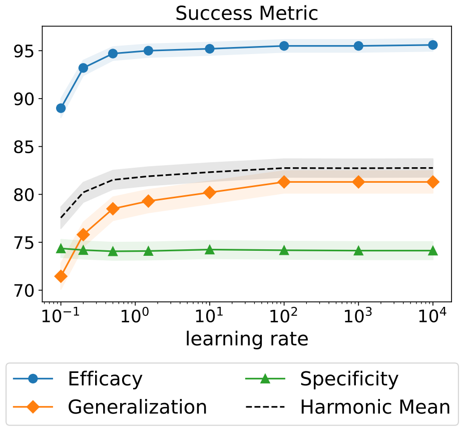
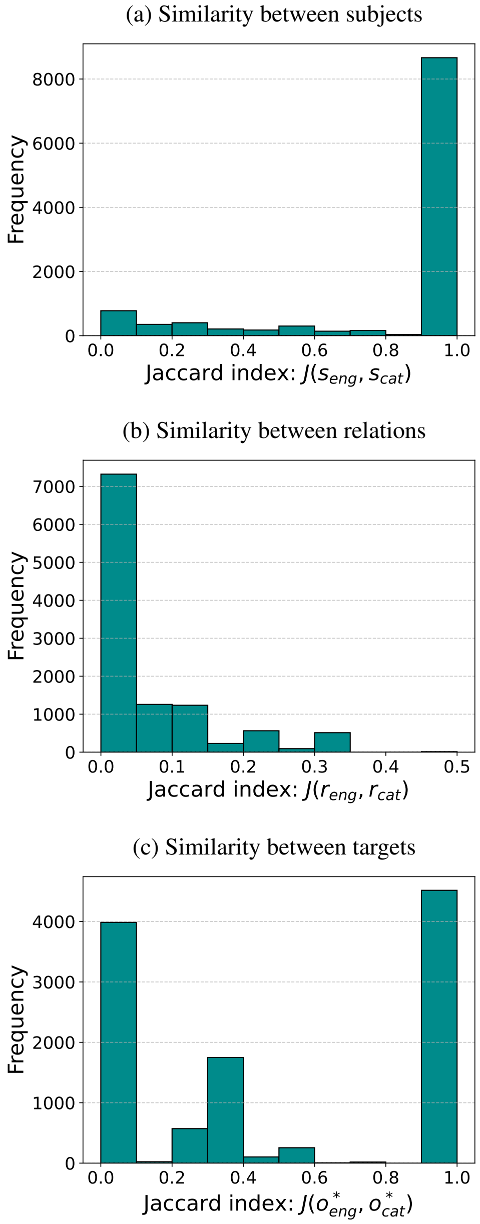


Figure 6: Results over training using PMET in Catalan and evaluating in Catalan with different values of the learning rate. Each solid line represents a different type of prompt used, and the dashed line represents the harmonic mean between them. The displayed areas correspond to a 68% confidence interval.

C Similarity between Setences

In Section 5.1, the connection between MEMIT's cross-lingual capability and subject tokenization becomes evident. Given that in some sections we also evaluate cross-lingual capacity without imposing restrictions on samples based on a specific Jaccard metric, we consider pertinent to depict the distribution of our subjects, relations, and targets $\langle s, r, o^* \rangle$ across both languages in Figure 7.


(a) Similarity between subjects

(b) Similarity between relations

(c) Similarity between targets

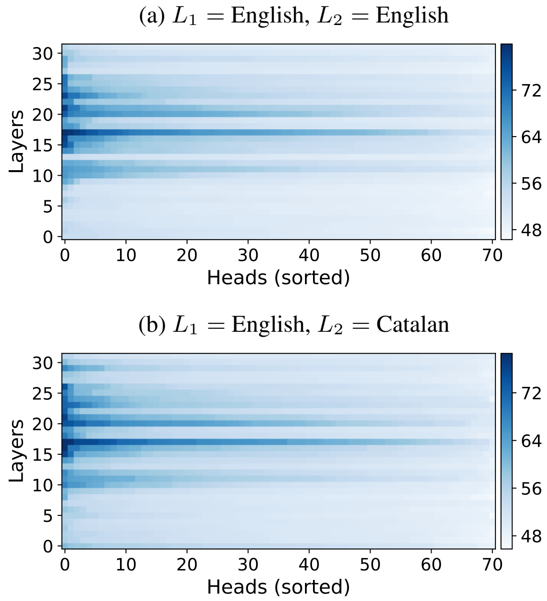
Figure 7: Visualization of the frequency distribution of similarities among subject, relation, and target tokens in the new English and Catalan CounterFact datasets.

D Implementation Details of Locating Relevant Heads

When incorporating information through a knowledge editor, not all intended factual associations are effectively inserted. To address this issue, our locating procedure incorporates an additional step. Following the insertion of factual associations using MEMIT in a designated language L_1, a refinement is made by selecting the associations that accurately predict the corresponding factual association in L_2. Subsequently, the locating procedure is applied to this refined subset of factual associations. This additional step enhances the top accuracy performance of certain aspects by approximately 8%.

E Relevance of Attention Heads

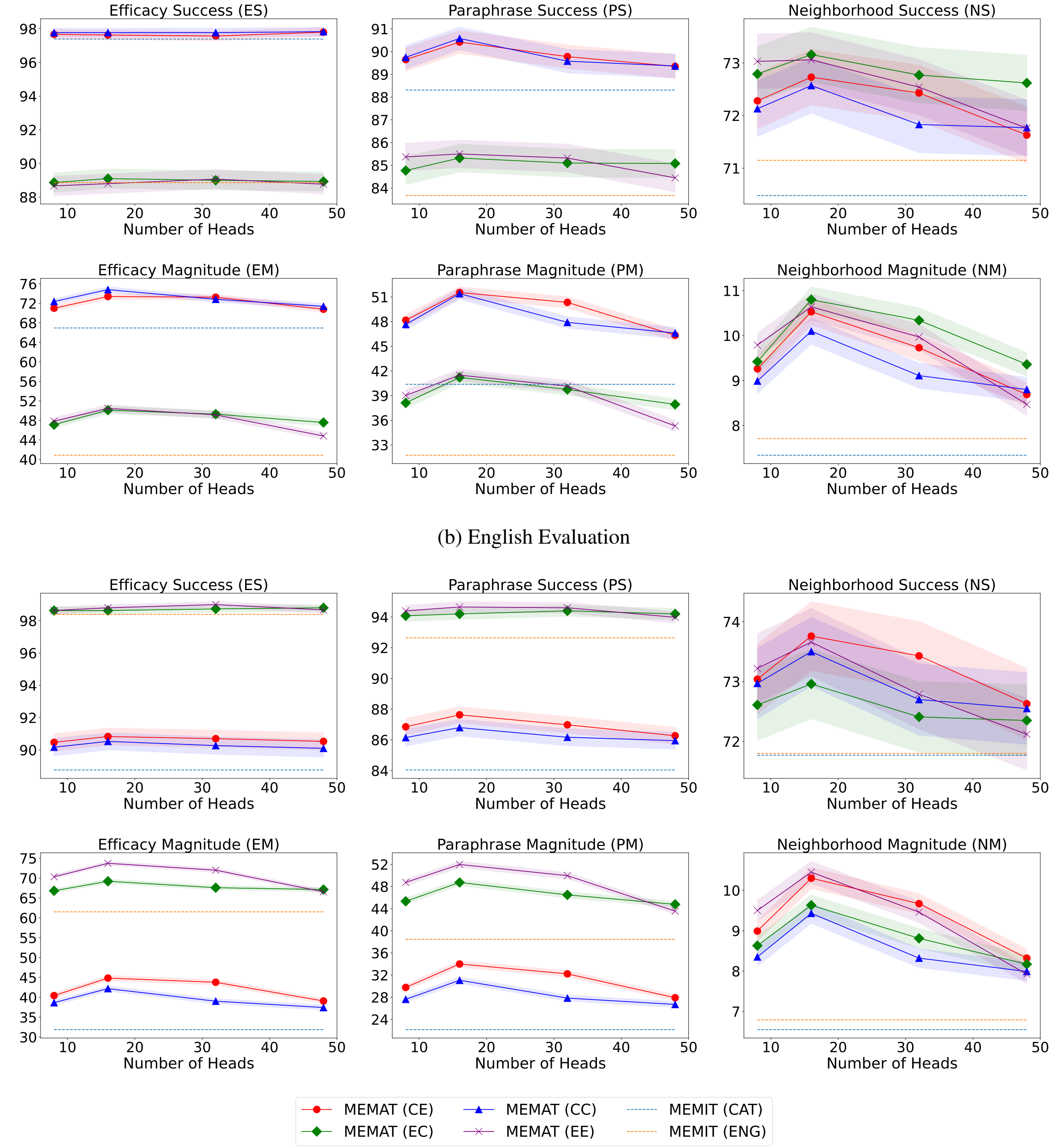
In Figure 8, we also explore the combinations of the languages L_1 and L_2 that were not previously represented.


(a) L_1 = English, L_2 = English

(b) L_1 = English, L_2 = Catalan

Figure 8: Accuracy on the validation set for all heads in all layers in Ǎguila-7B considering two combinations of L_1 and L_2. The performance peaks include 79.1% and 78.6%.

F Hyperparameter Search



(a) Catalan Evaluation

(b) English Evaluation

Figure 9: Illustration of all the metrics in Catalan (a) or English (b) evaluation when employing MEMAT for different number of heads ($K \in \{8, 16, 32, 48\}$), in 1,000 factual samples and considering all conceivable combinations of $L_1, L_2 \in \{\text{Catalan, English}\}$. The codification of L_1 and L_2 follows the format "L_1-L_2". The dashed lines represent the baseline performances when only applying a MEMIT training for Catalan or English. The displayed errors correspond to a 68% confidence interval.

G Tables of Results

Table 2 showcases the outcomes derived from the various knowledge editing methods investigated in this study, particularly focusing on previously unexplored combinations of training languages. Consistent with earlier findings, the incorporation of MEMAT yields a notable enhancement across the majority of evaluation metrics.

Furthermore, this performance boost extends to the accuracy metric, which assesses whether the most probable token is o^* for efficacy and paraphrase prompts or o^c for neighborhood prompts:

$$\begin{aligned} \text{EA} &:= \mathbb{E}\left[o^* = \arg\max_{\omega} \mathbb{P}[\omega|p] | p \in EP\right] \\ \text{PA} &:= \mathbb{E}\left[\mathbb{E}_{p\in PP}\left[o^* = \arg\max_{\omega} \mathbb{P}[\omega|p]\right]\right] \\ \text{NA} &:= \mathbb{E}\left[\mathbb{E}_{p\in NP}\left[o^c = \arg\max_{\omega} \mathbb{P}[\omega|p]\right]\right]. \end{aligned} \tag{13}$$

Method	English	Catalan	English	Catalan	English	Catalan
(Training Language(s))	ES	ES	PS	PS	NS	NS
Ǎguila-7B Baseline	26.5 (1.2)	25.5 (1.2)	30.4 (1.2)	31.6 (1.2)	73.9 (0.9)	72.2 (0.8)
PMET (ENG)	95.9 (1.2)	83.6 (2.3)	86.4 (1.8)	78.1 (2.4)	**74.0 (2.1)**	**74.3 (1.8)**
MEMIT (ENG)	98.4 (0.3)	88.9 (0.7)	92.6 (0.5)	83.7 (0.8)	71.8 (0.8)	71.2 (0.7)
MEMAT-16 (EE)	**99.0 (0.4)**	89.1 (1.1)	**94.6 (0.7)**	**85.3 (1.2)**	72.8 (1.1)	72.5 (1.0)
MEMAT-16 (EC)	98.7 (0.4)	89.0 (1.1)	94.4 (0.7)	85.1 (1.2)	72.4 (1.2)	72.8 (1.0)
MEMAT-16 (EE*)	98.6 (0.4)	89.3 (1.1)	94.4 (0.7)	85.0 (1.2)	73.4 (1.1)	73.0 (1.0)
MEMAT-16 (EC*)	98.5 (0.4)	**89.4 (1.1)**	93.7 (0.7)	84.8 (1.2)	72.6 (1.2)	72.6 (1.0)
	EM	EM	PM	PM	NM	NM
Ǎguila-7B Baseline	-6.7 (0.4)	-7.4 (0.5)	-5.5 (0.4)	-6.2 (0.5)	7.6 (0.3)	8.3 (0.4)
PMET (ENG)	68.3 (1.9)	38.7 (2.4)	36.0 (2.0)	29.4 (2.2)	8.8 (0.8)	7.9 (0.8)
MEMIT (ENG)	61.5 (0.7)	40.9 (0.9)	38.5 (0.7)	31.8 (0.8)	6.8 (0.3)	7.7 (0.3)
MEMAT-16 (EE)	**72.0 (1.0)**	49.1 (1.4)	**50.0 (1.1)**	**40.2 (1.4)**	9.5 (0.5)	10.0 (0.6)
MEMAT-16 (EC)	67.6 (1.1)	**49.3 (1.4)**	46.5 (1.1)	39.8 (1.4)	8.8 (0.5)	10.3 (0.6)
MEMAT-16 (EE*)	71.0 (1.0)	48.4 (1.5)	49.6 (1.1)	39.4 (1.4)	**9.9 (0.5)**	**10.4 (0.6)**
MEMAT-16 (EC*)	67.5 (1.1)	48.1 (1.4)	46.2 (1.1)	39.1 (1.4)	9.2 (0.5)	10.3 (0.6)
	EA	EA	PA	PA	NA	NA
Ǎguila-7B Baseline	0.3 (0.2)	0.8 (0.3)	0.3 (0.1)	1.3 (0.3)	9.8 (0.5)	13.0 (0.6)
PMET (ENG)	**87.8 (2.0)**	55.0 (3.1)	49.4 (2.6)	41.7 (2.9)	10.8 (1.1)	13.3 (1.3)
MEMIT (ENG)	78.8 (1.0)	57.7 (1.2)	52.0 (1.0)	46.1 (1.1)	9.6 (0.4)	12.3 (0.5)
MEMAT-16 (EE)	84.8 (1.3)	62.9 (1.7)	**62.2 (1.4)**	52.6 (1.7)	13.8 (0.7)	16.2 (0.8)
MEMAT-16 (EC)	83.2 (1.3)	**64.2 (1.7)**	60.7 (1.4)	**54.8 (1.7)**	13.7 (0.6)	**17.6 (0.8)**
MEMAT-16 (EE*)	85.2 (1.3)	63.2 (1.7)	62.1 (1.4)	53.4 (1.7)	**14.8 (0.7)**	16.7 (0.8)
MEMAT-16 (EC*)	82.8 (1.4)	63.3 (1.7)	59.7 (1.4)	53.8 (1.7)	14.0 (0.7)	16.7 (0.8)
PMET (CAT)	41.7 (3.1)	83.8 (2.3)	23.2 (2.2)	47.1 (3.0)	13.2 (1.3)	10.1 (1.1)
MEMIT (CAT)	45.2 (1.2)	82.6 (0.9)	33.4 (0.9)	55.7 (1.1)	8.8 (0.4)	12.7 (0.5)
MEMAT-16 (CC)	51.6 (1.8)	84.3 (1.3)	39.3 (1.4)	60.8 (1.7)	11.5 (0.6)	15.3 (0.8)
MEMAT-16 (CE)	**57.6 (1.8)**	84.6 (1.3)	**45.4 (1.5)**	**63.4 (1.6)**	**14.4 (0.7)**	16.6 (0.8)
MEMAT-16 (CC*)	54.5 (1.8)	**84.9** (1.3)	42.9 (1.5)	62.6 (1.6)	13.7 (0.7)	**16.8 (0.8)**
MEMAT-16 (CE*)	52.7 (1.8)	83.9 (1.3)	41.2 (1.5)	61.3 (1.7)	12.6 (0.6)	15.5 (0.8)

Table 2: Results of English and Catalan Efficacy, Generalization and Specificity prompts over the success, magnitude and accuracy metrics in both languages. Each row represents the experiments performed for the different knowledge editing methods when inserting 1,000 factual associations. The notation assigned to MEMAT-16 is (L_1-L_2), where the cases (L_1-L_2*) indicate the use of attention heads that were trained in a different set of factual triplets and which have been recycled in a new insertion of factual associations. The 95% confidence intervals are in parenthesis.

H Introduction of Both Languages

To incorporate factual associations for both languages using the Δ matrices defined in equation 5, we can employ two strategies:

- Optimize each factual association concurrently in English and Catalan, resulting in a single bilingual matrix, $\Delta_{eng+cat}$.
- Optimize two separate matrices for the same factual associations in English and Catalan independently, and then combine them: $\Delta_{eng} + \Delta_{cat}$.

Table 3 displays the experimental results for 1,000 samples, demonstrating that the second strategy—optimizing separate matrices before merging them—yields significantly superior outcomes. Additionally, the MEMAT approach for inserting both languages consistently outperforms MEMIT across most metrics. Comparing these findings with the monolingual insertion results in Tables 1 and 2, it is clear that while all neighborhood metrics decline, paraphrase metrics see substantial improvement. Furthermore, efficacy metrics are more balanced between the languages. These results indicate that inserting knowledge in both languages enhances the model's comprehension of factual concepts more effectively, though it also impacts unrelated knowledge to a greater extent.

I ITI performance

ITI proposes a method aimed at enhancing the accuracy of language models in the generation of truthful information. Firstly, the approach involves identifying the heads responsible for encoding pertinent information related to the concept of truth, which only differs with the locating procedure outlined in Section 5.2 in the monolingual framework and the dataset, which is TruthfulQA (Lin et al., 2022). Subsequently, an average of attention heads associated with the final token of truthful sentences is applied to the entire model. While ITI originally used a limited number of sentences, this Appendix study its robustness through an experiment comprising 1,000 samples.

The initial two stages of our experiment replicate the methodology outlined in Section 6. However, instead of optimizing the heads, we average the truthful samples and amplify the strength of the introductions by a factor of α. This approach yields the results shown in Figure 10. Although these results are subject to high statistical uncertainty, we find the outcomes from the optimization using 12 to be more favorable.

Method (Training Language(s))	English ES	Catalan ES	English PS	Catalan PS	English NS	Catalan NS
MEMIT (CAT+ENG)	93.4 (0.6)	92.2 (0.6)	83.7 (0.8)	84.8 (0.7)	**67.4 (0.7)**	**67.8 (0.8)**
MEMIT (CAT)+(ENG)	98.3 (0.5)	97.1 (0.6)	94.8 (0.6)	91.3 (0.9)	65.4 (1.2)	65.1 (1.1)
MEMAT (CAT)+(ENG)	**99.0 (0.4)**	**97.6 (0.6)**	**96.2 (0.6)**	**93.2 (0.8)**	67.2 (1.1)	66.6 (1.1)
	EM	EM	PM	PM	NM	NM
MEMIT (CAT+ENG)	36.7 (0.8)	28.3 (0.7)	23.5 (0.7)	18.0 (0.5)	5.3 (0.3)	4.8 (0.2)
MEMIT (CAT)+(ENG)	54.5 (1.1)	59.7 (1.2)	39.2 (1.0)	43.9 (1.3)	4.6 (0.4)	4.5 (0.5)
MEMAT (CAT)+(ENG)	**68.0 (1.1)**	**70.0 (1.2)**	**53.1 (1.1)**	**55.3 (1.4)**	**7.1 (0.5)**	**7.0 (0.6)**
	EA	EA	PA	PA	NA	NA
MEMIT (CAT+ENG)	55.2 (1.2)	42.1 (1.2)	37.0 (1.1)	27.9 (0.9)	10.0 (0.4)	7.4 (0.3)
MEMIT (CAT)+(ENG)	69.5 (1.6)	75.5 (1.5)	51.4 (1.5)	57.2 (1.7)	7.4 (0.4)	9.3 (0.6)
MEMAT (CAT)+(ENG)	**78.5 (1.5)**	**80.1 (1.4)**	**63.2 (1.4)**	**65.6 (1.6)**	**11.5 (0.6)**	**13.6 (0.7)**

Table 3: Results of English and Catalan Efficacy, Generalization and Specificity prompts over the success, magnitude and accuracy metrics in both languages. Each row represents experiments performed for different methods to insert factual knowledge in both languages. The 95% confidence intervals are in parenthesis.

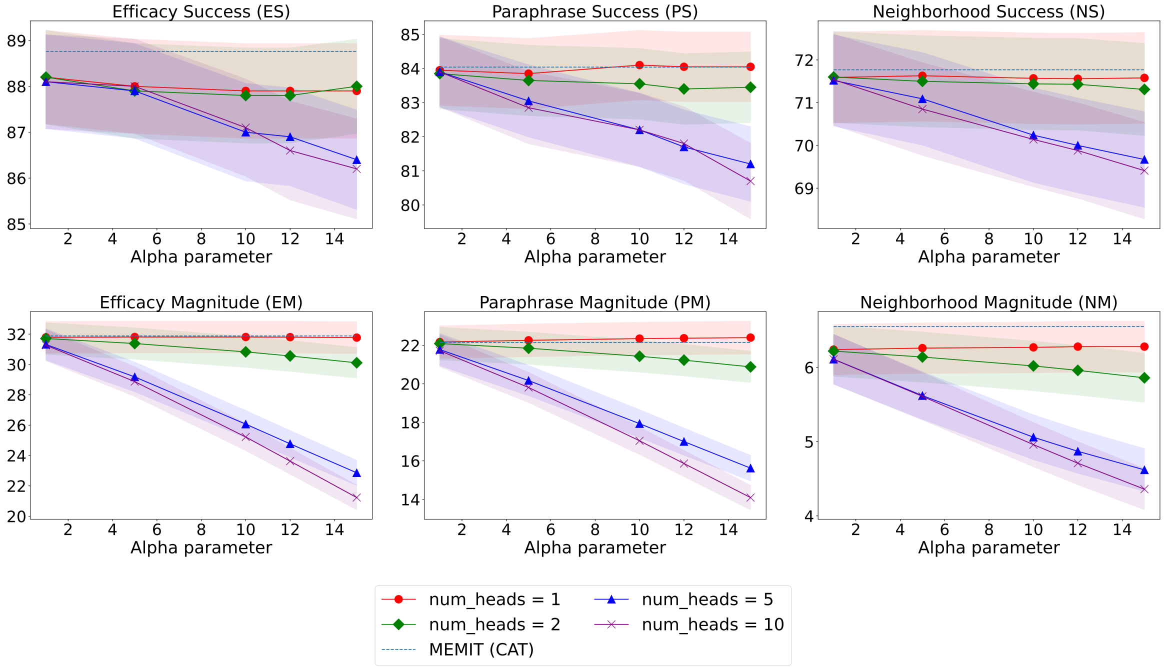


Figure 10: English Efficacy, Generalization and Specificity general metrics after applying ITI in the context of factual knowledge considering an original training in Catalan, and ITI applied in English. The interval of confidence considered is 68%.