Mass-Editing Memory with Attention in Transformers

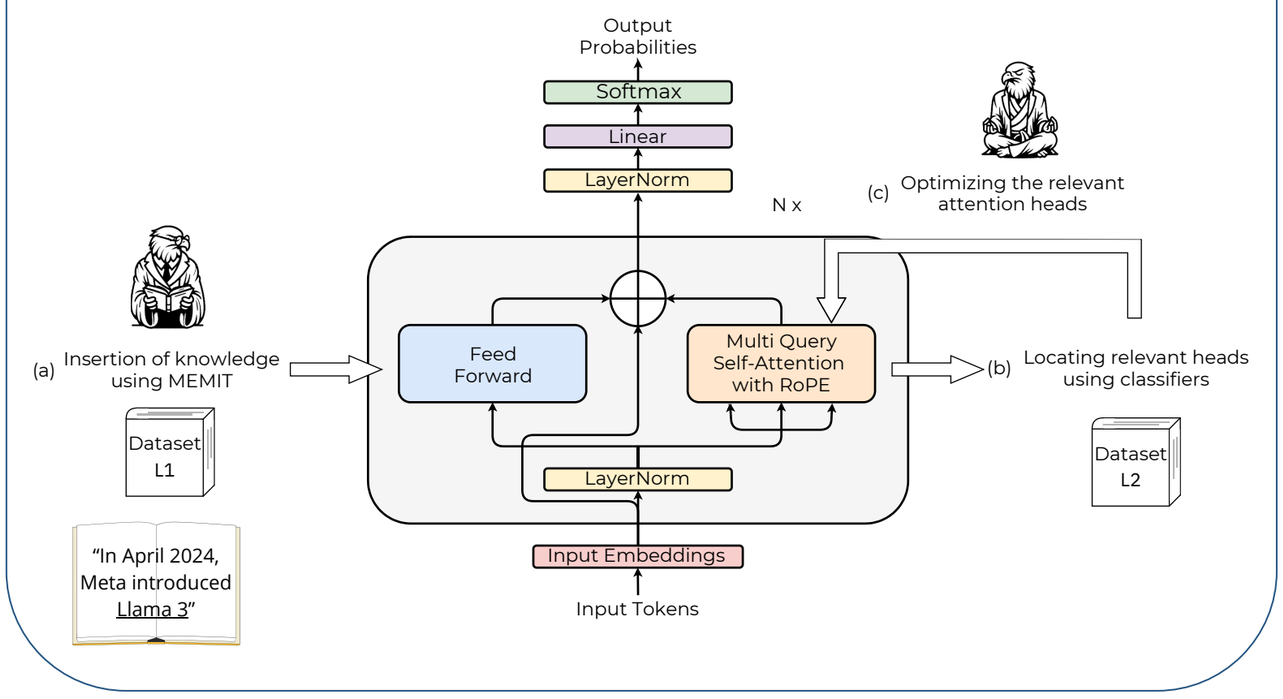
A cross lingual exploration of knowledge

Daniel Tamayo, Aitor Gonzalez-Agirre, Javier Hernando, Marta Villegas

Abstract

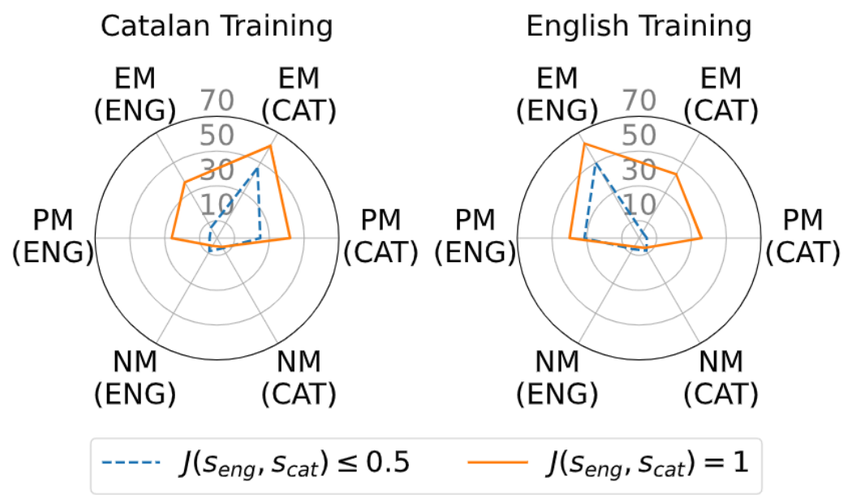
Recent research has explored methods for updating and modifying factual knowledge in large language models, often focusing on specific multi-layer perceptron blocks. This study expands on this work by examining the effectiveness of existing knowledge editing methods across languages and delving into the role of attention mechanisms in this process. Drawing from the insights gained, we propose MassEditing Memory with Attention in Transformers (MEMAT), a method that achieves significant improvements in all metrics while requiring minimal parameter modifications.

Overview of MEMAT



(a) Multilingual Challenge

Insertion of knowledge with MEMIT in one language and evaluation in both.

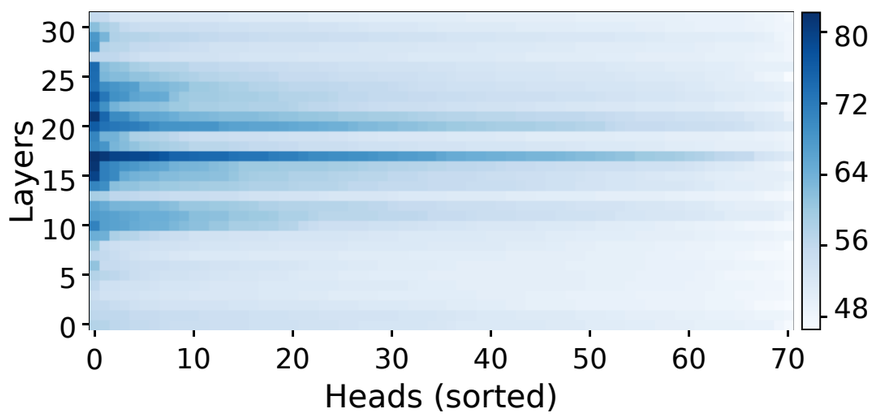


(b) Multilingual Capabilities

L1: Language in which we insert the knowledge with MEMIT.

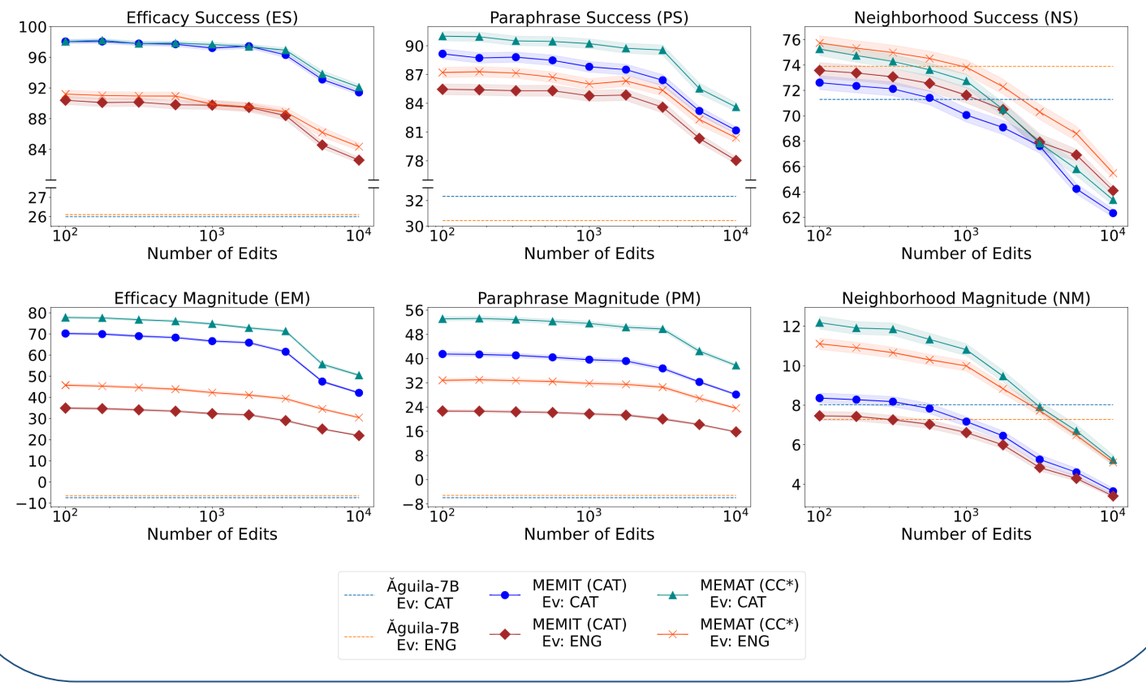
L2: Language with which we classify correct or incorrect sentences with classifiers in the attention heads.

L1 = Catalan, L2 = English



(c) Performance Enhancement

Insertion of knowledge in Catalan and evaluation in English and Catalan.



Conclusions

- 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.
- There is evidence that attention heads encode information about the truthfulness of factual associations in a language-independent manner.
- The methods explored struggle when subjects do not share the same tokenization between languages, which may yield different results for different pairs of languages.

This work has been promoted and financed by the Generalitat de Catalunya through the Aina project.

Este proyecto con referencia 2022/TL22/00215337 está financiado por el Ministerio para la Transformación Digital y de la Función Pública y por el Plan de Recuperación, Transformación y Resiliencia -Financiado por la Unión Europea - NextGenerationEU.