On the Representations of Entities in Auto-regressive Large Language Models

Victor Morand[1] Josiane Mothe[2] Benjamin Piwowarski[1]

[1] ISIR, Sorbonne Université [2] IRIT, Université de Toulouse

Introduction

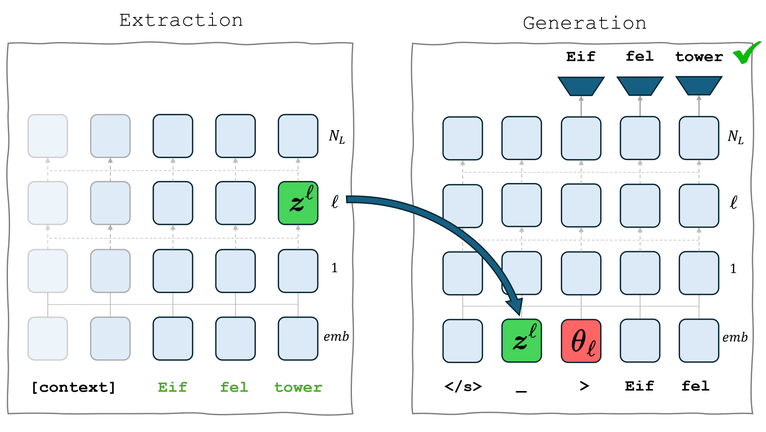
On the vast web of knowledge, entities are the atoms: the fundamental units that everything else builds upon. Although some clues emerge from mechanistic interpretability, how auto-regressive LLMs encode and retrieve these entities remains a mystery.

- Previous work **[3,4]** hypothesize that best entity representation are found in *middle-layer representations of the last token of their mention*. Following them, we posit that LLMs compute layer-agnostic entity representations that can be isolated and manipulated. Our primary objective is to establish a direct association between these internal representations and named entities.
- To evaluate this association, we measure how accurately the corresponding mention can be generated from the representation at hand.

Research Questions

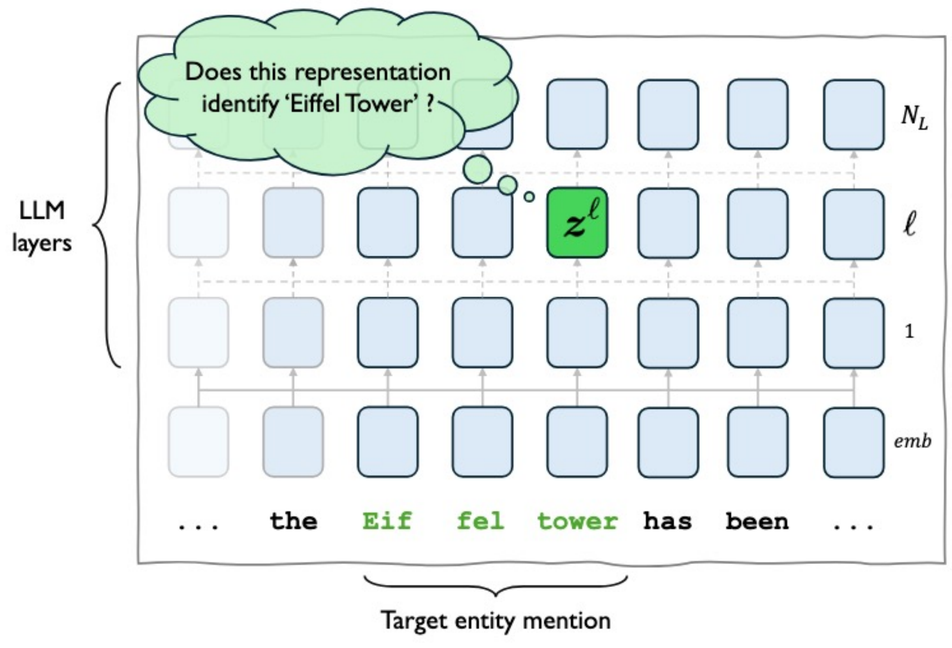
- **RQ 1** : How well can entity mentions be decoded from their representation?
- **RQ 2** : Can we find better representations than the last token representations from LLMs?
- **RQ 3** : Does the structure of the entity representation space capture (relational) knowledge?

Method – Mention Reconstruction with learned Task Vectors



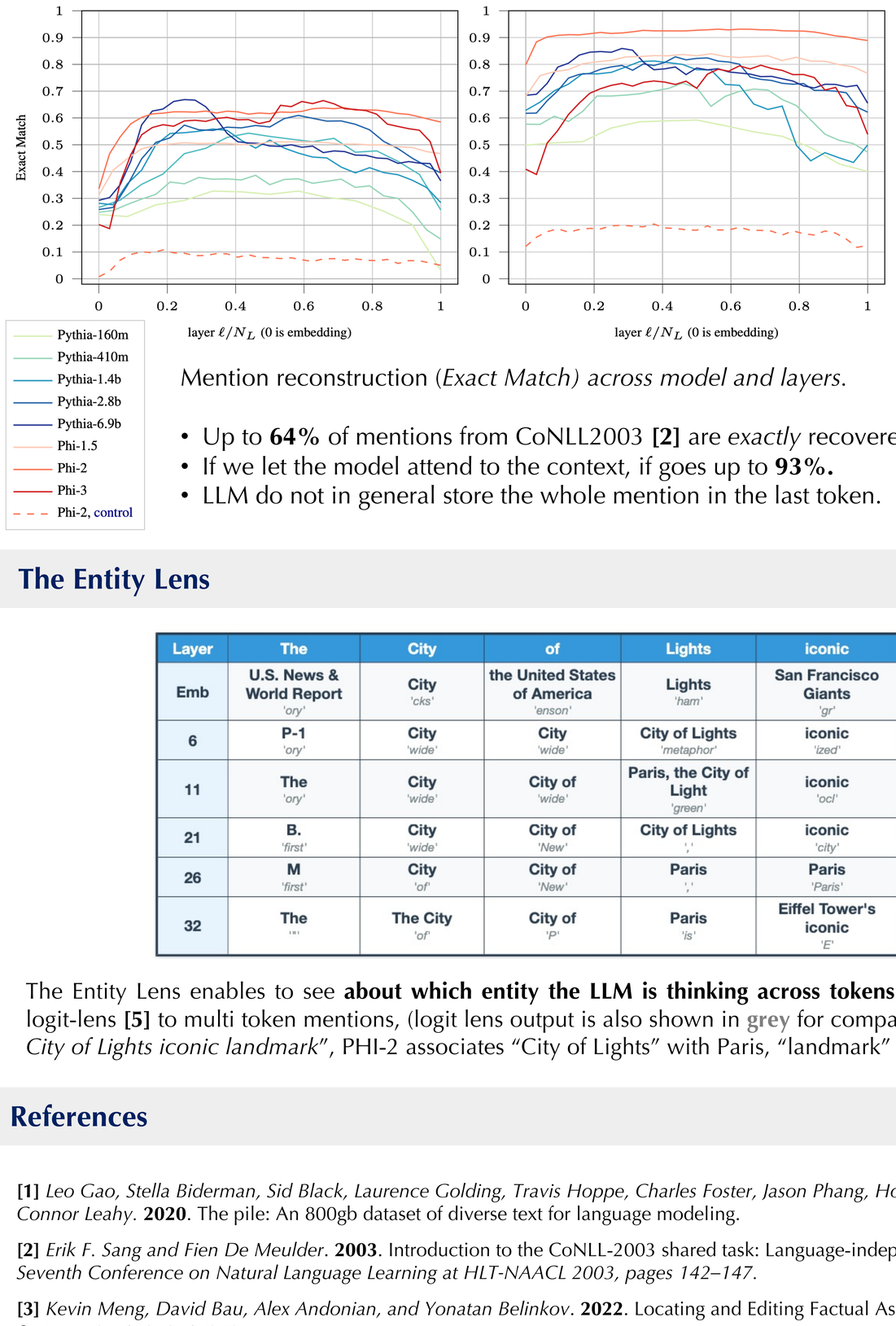
Uncontextual entity mention decoding: The entity representation $\mathbf{z}^\ell$ at layer ℓ is extracted in context (left, green). The LLM is then prompted to to generate the mention from $\mathbf{z}^\ell$ only, using a learned **task vector** θ_ℓ .

We also experiment with a **Contextual entity mention decoding**, where the LLM can attend to the context when generating the mention.



Results

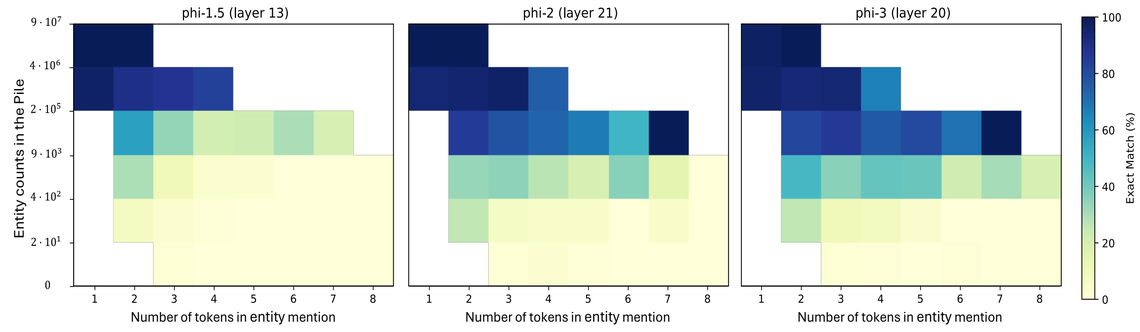
Mention Reconstruction performance across Models and Layers



Mention reconstruction (*Exact Match) across model and layers*.

- Up to **64%** of mentions from CoNLL2003 **[2]** are *exactly* recovered.
- If we let the model attend to the context, if goes up to **93%.**
- LLM do not in general store the whole mention in the last token.

Common Entities are almost part of LLMs vocabulary



Uncontextual mention generation performance is higher for more frequent entities. Performance is analysed by entity length and mention frequency in *the Pile* **[1]**. For each model, we chose the layer with best exact match on the test set. Empty cells indicate fewer than five samples.

- Very frequent entities are exactly reconstructed, *regardless of their length*.
- Mention Frequency has greater impact on reconstruction than token length.
- Bigger models "remember" more entities.

The Entity Lens

Layer	The	City	of	Lights	iconic	landmark
Emb	U.S. News & World Report *'ory'*	City *'cks'*	the United States of America *'enson'*	Lights *'ham'*	San Francisco Giants *'gr'*	Hague *'olla'*
6	P-1 *'ory'*	City *'wide'*	City *'wide'*	City of Lights *'metaphor'*	iconic *'ized'*	iconic landmark *'ry'*
11	The *'ory'*	City *'wide'*	City of *'wide'*	Paris, the City of Light *'green'*	iconic *'ocl'*	iconic landmark *'status'*
21	B. *'first'*	City *'wide'*	City of *'New'*	City of Lights *','*	iconic *'city'*	landmark *','*
26	M *'first'*	City *'of'*	City of *'New'*	Paris *','*	Paris *'Paris'*	landmark *','*
32	The *'"'*	The City *'of'*	City of *'P'*	Paris *'is'*	Eiffel Tower's iconic *'E'*	Eiffel Tower *','*

The Entity Lens enables to see **about which entity the LLM is thinking across tokens and layers**, extending the logit-lens **[5]** to multi token mentions, (logit lens output is also shown in **grey** for comparison). When reading "*The City of Lights iconic landmark*", PHI-2 associates "City of Lights" with Paris, "landmark" with the Eiffel Tower.

Conclusions

1. LLMs do compute layer-agnostic entity representations that can be isolated and manipulated.
2. Common entity mentions are (almost) part of the Vocabulary. They can be fully recovered from middle-layer representations of their last token. **Still,** uncommon mentions aren't fully encoded this way but rather retrieved from the context.
3. By successfully learning "Tasks Vectors" steering the model to reconstruct the mention, we uncover new evidence that LLMs form dedicated internal circuits to represent and manipulate multi-token entities.

References

[1] *Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy*. **2020**. The pile: An 800gb dataset of diverse text for language modeling.

[2] *Erik F. Sang and Fien De Meulder*. **2003**. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, pages 142–147*.

[3] *Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov*. **2022**. Locating and Editing Factual Associations in GPT. *Advances in Neural Information Processing Systems, 35:17359–17372*.

[4] *Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson*. **2023**. Dissecting recall of factual associations in auto-regressive language models. *In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 12216–12235, Singapore*.

[5] *nostalgebraist*. **2020**. Interpreting GPT: The logit lens. *LessWrong*.

EMNLP 2025 Suzhou, China 中国苏州

morand [@] isir.upmc.fr