Investigating Knowledge Closure of Large Language Models via Token Embeddings

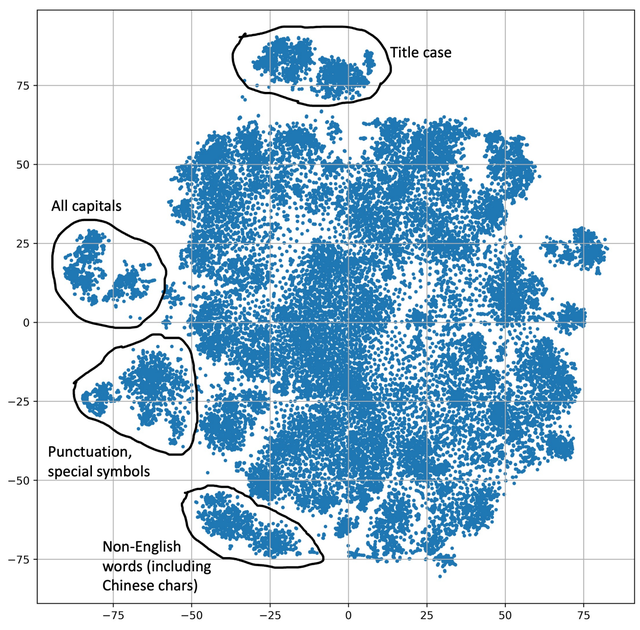
Sourya Dey (sourya@galois.com)*, Michael Robinson†, Taisa Kushner*, Andrew Lauziere*, Cait Burgess*

Research conducted by *Galois and †American University for the DARPA Emergent Risks program

Broad Goal: *Our research characterizes the behavior and potential risks of large language models (LLMs) via: a) analyzing their internal parameters, b) investigating the effects of structured prompting.*

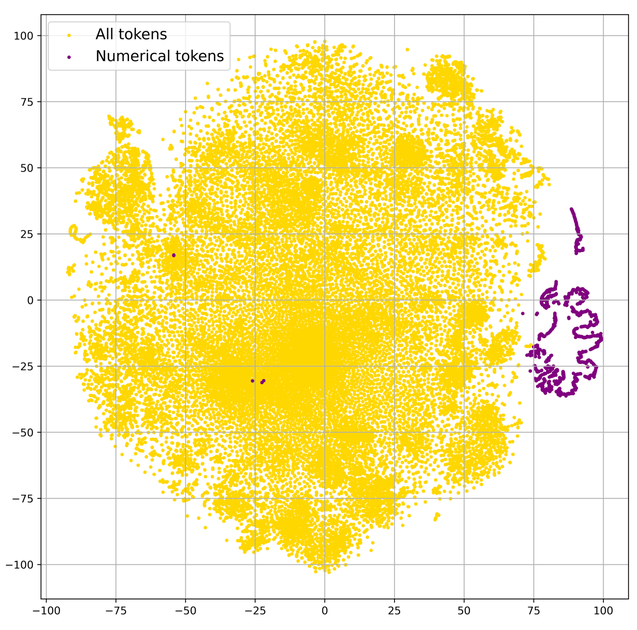
Token Embeddings

Tokens in the vocabulary of a LLM are internally stored as embedding vectors in a high-dimensional space. Visualizing this space in 2D via PCA and t-SNE uncovers clusters of semantically or syntactically related tokens.



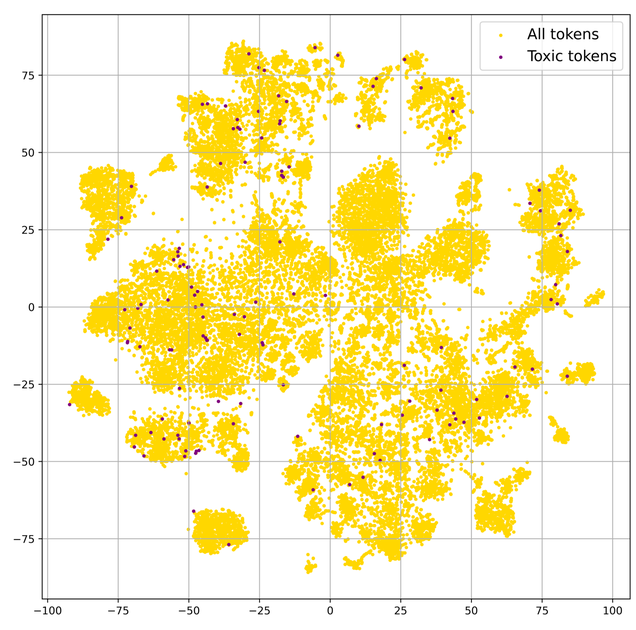
Llemma-7B: A LLM based on CodeLlama.

Syntactically related tokens such as *capitalized words* form clusters.



Pythia-6.9B: One of a family of LLMs based on the GPT-Neo (GPT-3) architecture.

Numerical tokens cluster in one area.

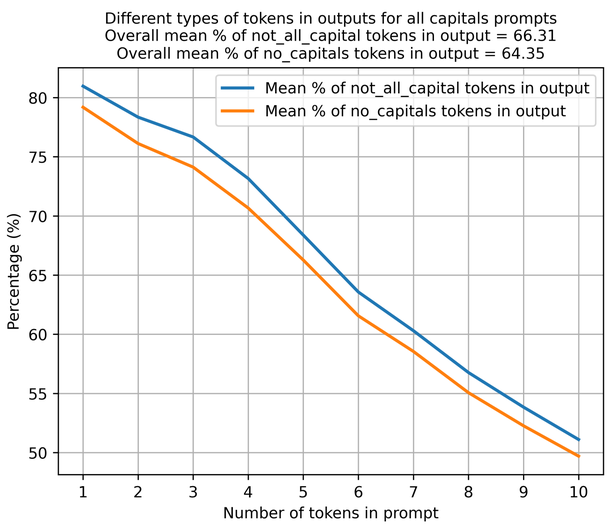


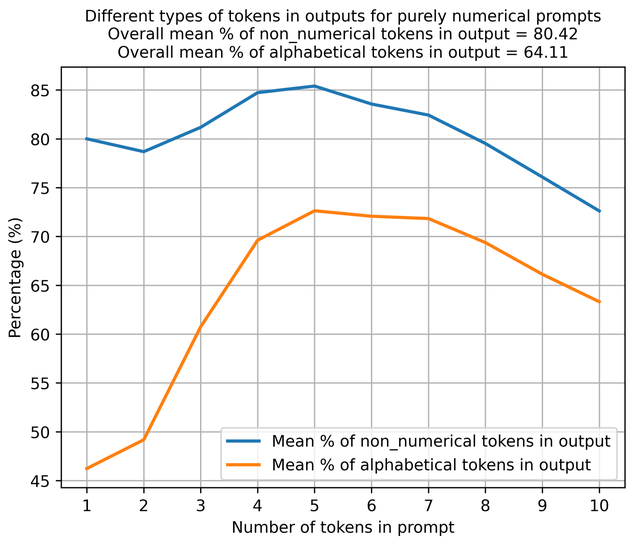
Mistral-7B: A LLM to compete with Llama.

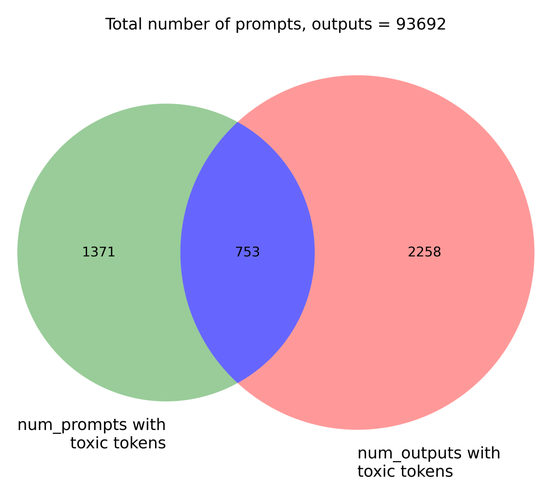
Unlike other related tokens, *tokens with toxic language* in them are *not* clustered.

Knowledge Closure

Can LLMs be used to produce knowledge that is unsafe, e.g. controversial or dangerous? Alternatively, can LLMs break free of existing syntactic / semantic / embedding domains of knowledge present in their prompts and produce new knowledge?







Prompting Llemma-7B (left) with only capital tokens produces many output tokens that are not capitalized; likewise, prompting Pythia-6.9B (right) with only numerical tokens produces many non-numeric output tokens. However, the model's tendency to 'deviate from the script' reduces when the prompt has more tokens, i.e. a longer prompt in one knowledge domain makes the model more likely to stay in that domain.

Prompting LLMs with random strings of tokens yields a high chance of tokens with toxic language occurring in the output when they are not present in the prompt. I.e., toxic tokens have high reachability within the embedding space. This can be problematic in terms of preventing models from producing harmful language.

Results shown for Mistral-7B, similar trends seen for other models.

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited).

galois