Internship Proposal

Title: The Effect of Prototypical Reasoning in Large Language Models
Keywords Large Language Models, Inference optimization, bias and fairness in NLP
Supervisor: Pr. Julien Velcin, Ecole Centrale de Lyon, LIRIS
Location: Ecole Centrale de Lyon, 36 Av. Guy de Collongue, 69130 Écully
Duration: 4 months (ideally April - July)
Stipend: About 650 euros / month

Context

Large Language Models (LLMs) have shown unprecedented capabilities in reasoning, classification, and text generation. With their growing scale, recent efforts have focused on reducing their computational cost using techniques such as knowledge distillation, quantization, and prototype-based reasoning. These methods promise faster inference and reduced hardware consumption. However, prior studies have shown that LLMs trained on biased data may reproduce harmful stereotypes, particularly when their internal representations are compressed or simplified (Bolukbasi et al., 2016). Representational biases can be exacerbated by inference speed-up mechanisms, potentially leading to subtle but socially significant harms.

This internship takes place in the context of the ANR project DIKé (`https://www.anr-dike.fr`).

General Problematic

Do techniques used to accelerate LLM inference, especially prototype-based reasoning, amplify the emergence of stereotypical representations?

While most bias evaluations focus on generation tasks, stereotypical shortcuts may also arise in classification and information retrieval, domains where prototypes can be used to improve performance without drawing attention to how they shape decision boundaries.

This project proposes to test whether speed-up techniques lead to decisions that are *efficient but stereotyped*, and to draw parallels with Walter Lippmann's theory of stereotypes as oversimplified mental images used to reduce cognitive load.

Objectives

The main objective is to evaluate and explain how prototype-based acceleration mechanisms affect different interesting metrics such as:

- inference speed
- accuracy on classification or information retrieval tasks
- bias metrics, for instance based on targeted group annotations

It would be very interesting to demonstrate to what extent accelerated inference results from leveraging biased prototypes.

Plan

1. **Literature Review** This review is broadly divided into three main directions. First, you should be more familiar with the theories of categorization (cf. work of E. Rosch or W. Lippmann). Second, you have to understand when and how the LLMs can be biased (see Caliskan et al., 2017; May et al., 2019). Third, you have to study the various techniques that are used for accelerating the inference of LLMs (Liu et al., 2025) and focus on prototype-based techniques (Proskurina et al., 2025).
2. **Corpus Selection** Based on the literature, you have to select a set of datasets that are relevant for the study. It can be based, for instance, on datasets related to hate speech detection, such as SBIC (Sap et al., 2020), which contains bias-related annotations, or IHC (ElSherief et al., 2021).
3. **Experimental Framework** The experiment setup will compare different LLMs that support the acceleration mechanisms. Those experiments will use the complementary metrics mentioned above (accuracy, inference time, bias). You will investigate what kind of prototypes has been used and analyze if those high-performing prototypes resemble stereotypical generalizations. This can lead to a discussion on implications for fairness in AI deployment.

Profile

Master 2 or last year engineer student, with experience in the usage of Pytorch and LLM pipelines (huggingface, langchain)

Contact

Please send your application to: julien.velcin@ec-lyon.fr

- CV
- Motivation letter
- Recent transcripts

References

Bolukbasi, T. et al. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. NIPS

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science

Crawford, K. (2017). The Trouble with Bias. NIPS Keynote

ElSherief, M., Ziems, C., Muchlinski, D., Anupindi, V., Seybolt, J., De Choudhury, M., & Yang, D. (2021). Latent hatred: A benchmark for understanding implicit hate speech. EMNLP

Liu, Y. et al. (2025). Efficient Inference for Large Reasoning Models: A Survey. arXiv:2503.23077

May, C. et al. (2019). On Measuring Social Bias in Sentence Encoders. NAACL

Proskurina I., M.A. Carpentier, J. Velcin (2025). HatePrototypes: Interpretable and Transferable Representations for Implicit and Explicit Hate Speech Detection, https://arxiv.org/abs/2511.06391

Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., & Choi, Y. (2020, July). Social bias frames: Reasoning about social and power implications of language. ACL