MILES
Machine Intelligence and Learning Systems

Reverse-engineering LLMs: making the most of the context

- **Keywords:** Large Language Models, Prompt, Natural Language Processing, Explainability.
- **Duration:** 4 à 6 mois.
- **Supervisors:**
  Alexandre Allauzen: `alexandre.allauzen@dauphine.fr`
  Florian Le Bronnec: `florian.le-bronnec@dauphine.psl.eu`
- **Place:** MILES, LAMSADE, Université Paris Dauphine (mainly at PariSanteCampus).

Context: Large Language Models (LLMs) display notable variations in performance based on the nature of in-context prompts. This performance discrepancy is particularly evident when tasks are vaguely described, resulting in poorer outcomes compared to scenarios where prompts include specific examples, called few-shots in-context learning [1]. Despite the significance of prompt influence on LLMs' performance, this aspect remains relatively unexplored.

Goals: The primary goal of this research internship is to delve into the factors that contribute to the performance variations of LLMs based on in-context prompts. The aim is to identify the key elements essential for LLMs to exhibit robust performance. The project also involves conducting an extensive literature review to glean insights from existing methodologies and findings. Building on this knowledge, the objective is to optimize prompts for LLMs, drawing inspiration from successful methods such as AutoPrompt [3] or methods only used in images [2]. Additionally, the investigation will explore the integration of examples in prompts for zero-shot learning and assess the feasibility of inferring this information using similar techniques. The overarching objective is to engage in advanced prompt research, utilizing techniques like gradient analysis to unravel the inner workings of LLMs.

Schedule:

1. **Literature Review:** Conduct a thorough review of existing literature on LLMs, focusing on the influence of prompts on performance. Summarize key findings and methodologies employed in previous research.
2. **Initial Experiments:** Replicate experiments akin to AutoPrompt to determine optimal prompts for LLMs. Evaluate the impact of various prompt types on the performance of LLMs.
3. **Advanced Prompt Research:** Delve into the inner workings of LLMs through gradient-based analyses. Gain insights into how gradients can be harnessed to understand and optimize prompts for improved LLM performance.
4. **Incorporating Examples:** Expand the research to include the need for examples in prompts, especially in zero-shot learning scenarios. Investigate the potential inference of example requirements by making the link with above methods.

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[2] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, October 2019.

[3] Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. AutoPrompt: Eliciting knowledge from language models with automatically generated prompts. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2020.