Stochastic Chameleons : Irrelevant Context Hallucinations Reveal Class-Based (Mis)Generalization in LLMs

Ziling Cheng[1,2*] **Meng Cao**[1,2*] **Marc-Antoine Rondeau**[1] **Jackie Chi Kit Cheung**[1,2,3]
[1]Mila – Quebec Artificial Intelligence Institute
[2]McGill University [3]Canada CIFAR AI Chair
{ziling.cheng, meng.cao}@mail.mcgill.ca, {ma.rondeau, cheungja}@mila.quebec

Abstract

The widespread success of large language models (LLMs) on NLP benchmarks has been accompanied by concerns that LLMs function primarily as stochastic parrots that reproduce texts similar to what they saw during pre-training, often erroneously. But what is the nature of their errors, and do these errors exhibit any regularities? In this work, we examine irrelevant context hallucinations, in which models integrate misleading contextual cues into their predictions. Through behavioral analysis, we show that these errors result from a structured yet flawed mechanism that we term *class-based (mis)generalization*, in which models combine abstract class cues with features extracted from the query or context to derive answers. Furthermore, mechanistic interpretability experiments on Llama-3, Mistral, and Pythia across 39 factual recall relation types reveal that this behavior is reflected in the model's internal computations: (i) abstract class representations are constructed in lower layers before being refined into specific answers in higher layers, (ii) feature selection is governed by two competing circuits — one prioritizing direct query-based reasoning, the other incorporating contextual cues — whose relative influences determine the final output. Our findings provide a more nuanced perspective on the stochastic parrot argument: through form-based training, LLMs can exhibit generalization leveraging abstractions, albeit in unreliable ways based on contextual cues — what we term *stochastic chameleons*.[1]

[*] Equal contribution.

[1] Code available at: https://github.com/ziling-cheng/Irrelevant-Context-Hallucination.