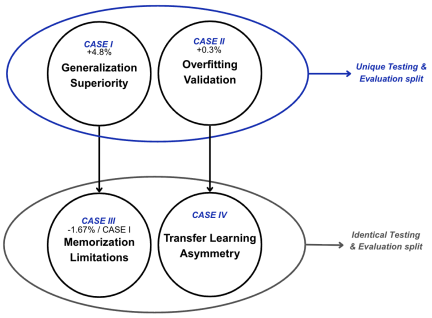
Finding Memo: The Hidden Influence of Memorization in Large Language Models' Performance – A Critical Analysis of Benchmark Evaluation

Large Language Models have achieved remarkable results on symbolic and mathematical reasoning tasks. However, current LLMs evaluation methods primarily assess overall performance but cannot reliably disentangle genuine generalization from mere memorization, meaning high scores may reflect data recall rather than reasoning ability. This ambiguity persists due to the absence of a robust, concise metric capable of disentangling these two phenomena. Secondly, the relation between model performance that varies with the familiarity of patterns in the training data suggests a dependence on memorization rather than true reasoning. Lastly, the influence of entity exposure frequency during training remains poorly understood, and its effect on performance can't be quantified.

This research addresses this gap by introducing a new benchmark dataset, ALSA-5K, that stands for: Arithmetic Learning and Symbolic Memorization Assessment, a purpose-built collection of math questions spanning 50 distinct real-world domains. The primary utility of ALSA-5K lies in its ability to measure the extent to which performance improvements are driven by memorization versus generalization by providing deeper insights into model behaviour under four main scenarios. The initial version of ALSA-5K includes variable naming for each name instance, enabling variation in name distributions that allow for precise diagnosis of memorization dynamics independent of reasoning ability. Using the Qwen-2.5-1.5B model, a series of fine-tuning experiments systematically probe different memorization and generalization conditions. Evaluation with a balanced-exact-match accuracy metric reveals that performance improvements are most pronounced in scenarios favouring memorization – such as repeated or identical name exposure, while gains in generalization-oriented settings remain limited. a) CASE I – Train on Unique Names, Test on Unique Names: Measures true generalization by ensuring the model never sees repeated names, eliminating memorization cues. b) CASE II – Train on Identical Names, Test on Unique Names: Assesses the overfitting/generalization gap, revealing how repeated exposure to the same names during training affects performance on novel entities. c) CASE III – Train on Identical Names, Test on Identical Names: Quantifies maximum memorization benefit, showing how repeated name exposure can inflate performance without reflecting real reasoning ability. d) CASE IV – Train on Unique Names, Test on Identical Names: Tests robustness under repetition, evaluating whether learning from diverse names allows the model to adapt to repetitive testing conditions. This structured evaluation highlights the hidden influence of memorization on LLM performance.



The ALSA-5K experiments provide strong evidence that symbolic reasoning in LLMs is driven by generalization at the highest rate. Nonetheless, memorization demonstrated performance levels that nearly matched – and at times exceeded – those of generalization. Across CASE I–IV, diverse training consistently outperformed memorization-based methods, with CASE I achieving the highest gain (+4.8%) and memorization showing clear performance ceilings at CASE III. Results also reveal transfer learning asymmetry, where models trained on diverse data adapt better to repetitive setups than the reverse, and validate the overfitting hypothesis with minimal gains in CASE II. Collectively, these findings underscore the necessity of varied training for achieving robust symbolic reasoning. Aligning with our hypothesis, Frequency-biased tests further demonstrate a strong correlation between exposure frequency and model accuracy, confirming the model's reliance on memorized patterns sometimes. Moreover, this suggests that memorization can be disentangled from reasoning ability through targeted benchmarks and probing techniques that lead to more accurate and trustworthy LLM evaluation.