Dual Operating Modes of In-Context Learning

Ziqian Lin & Kangwook Lee
THE UNIVERSITY of WISCONSIN MADISON

LLM's In-Context Learning (ICL)
= Learning a new task + Recognizing a known task

Llama3-70b-instruct, Claude-3-5-sonnet-20240620, gpt-4o-2024-05-13

Given the following examples,
find the value of X without any explanation.

hello => 5 whether => 7 black => 5
forecast => 8 amazing => 7 apple => 5
banana => X

LLMs' answer: 6 ✔

Hint: It's related to the exponential function.
2 @ 8 => 64 3 @ 5 => 125 2 @ 5 => 25
4 @ 3 => X

LLMs' answer: 64 ✘

Banana => **Black** Apple => **Gray**
Watermelon => **White** Cherry => **Purple**
Strawberry =>

LLMs' answer: **Red** ?

gpt-3.5-turbo-instruct

1 @ 5 = 7
2 @ 10 = 13
4 @ 4 = 9
5 @ 8 = { 13 (93.05%) / 14 (0.59%) } ✘

1 @ 5 = 7
2 @ 10 = 13
4 @ 4 = 9
...
+30 demos
...
5 @ 8 = { 13 (11.98%) / 14 (66.44%) } ✔

gpt-4o-2024-05-13

6 @ 5 => 15625
6 @ 3 => 729
3 @ 4 => 64
...
+ 75 demos
...
2 @ 5 => 25
4 @ 3 => 81
2 @ 3 => 9
5 @ 2 => ?
Let's think step by step

LLMs' answer:
Let's start by identifying the pattern in the operations given.
...
Based on these calculations, ...
5@2 = 25. ✘

Q. Can you classify each of the cases above into learning-dominant and retrieval-dominant?

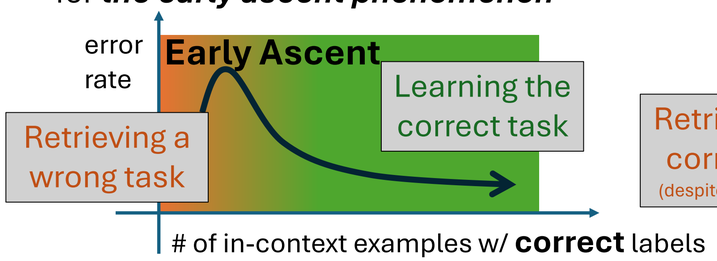
Key Research Questions

Q1. Can we develop a mathematical model that can explain the **dual operating modes**? (1)
Q2. Can we explain the real LLMs' phenomena with our model & analysis? (2) (3)

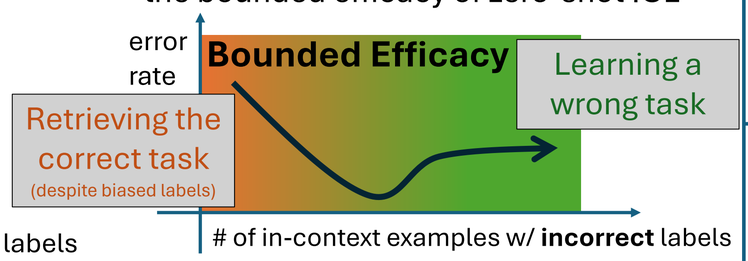
Our Contributions

(1) We rigorously explain the dual operating modes of ICL via a mathematical model

(2) We provide the first theoretical explanation for ***the early ascent phenomenon***

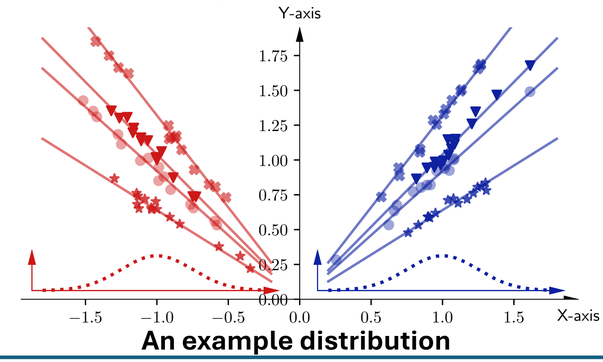


(3) We predict and empirically confirm the bounded efficacy of zero-shot ICL

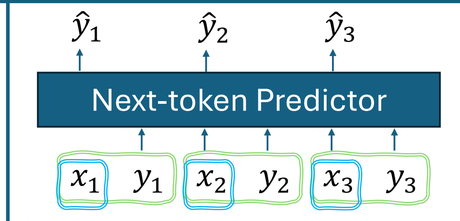


(1) Our model: An extension of [1, 2]:

(A) Linear functions ($y = w^T x$)
(B) x distribution & w distribution are **dependent**
(C) TF is perfectly pretrained with MSE, i.e., MMSE



An example distribution

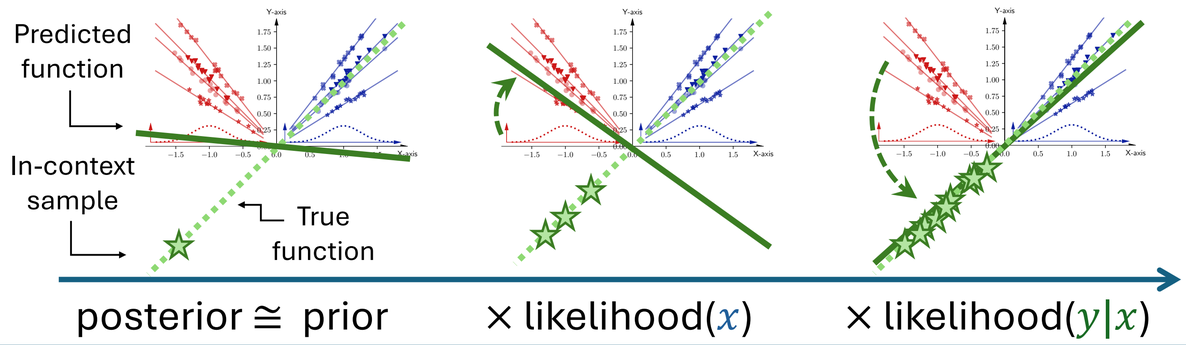


$$y^*_{\text{test}} = E[w|\text{prompt}]^T x_{\text{test}}$$

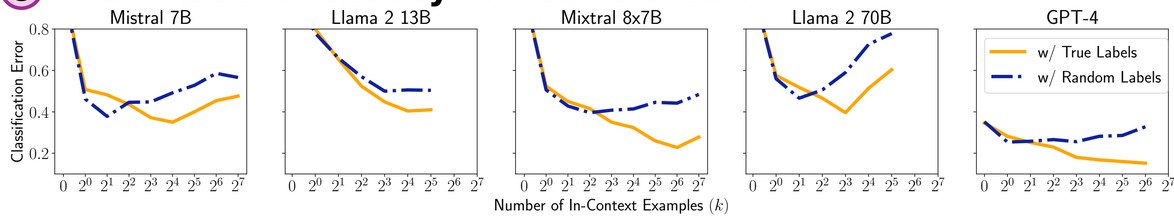
Many in-context samples
➔ Sufficient information about y/x
➔ The posterior mean converges to w
➔ Learning a new task

Few in-context samples
➔ Information about x is dominant
➔ A correct task retrieved under some conditions

(2) The *early ascent* phenomenon: incorrect retrieval + learning



(3) The bounded efficacy of zero-shot ICL



Related work

[1] What can Transformers learn in-context? A case study of simple function classes. Garg et al. (2022)
[2] The effects of pretraining task diversity on in-context learning of ridge regression. Raventos et al. (2023)
[3] What in-context learning "learns" in-context: Disentangling task recognition and task learning. Pan et al. (2023)
[4] Z-ICL: Zero-shot in-context learning with pseudo-demonstrations. Lyu et al. (2023)

Acknowledgement

The work of Kangwook Lee is supported in part by NSF CAREER Award CCF-2339978, Amazon Research Award, and a grant from FuriosaAI.