Fine-tuning vs. In-context Learning in Large Language Models

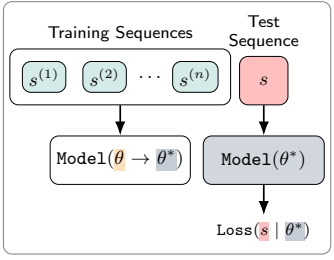
A Formal Language Learning Perspective

Bishwamittra Ghosh, Soumi Das, Till Speicher, Qinyuan Wu, Mohammad Aflah Khan, Deepak Garg, Krishna P. Gummadi, Evimaria Terzi

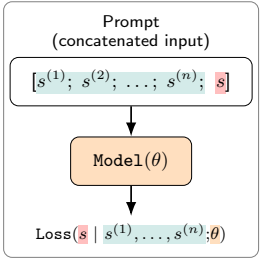
Max Planck Institute for Software Systems, Germany
Boston University, USA

Two Learning Modes

Large Language Models (LLMs) learn in two fundamental modes



Fine-Tuning



In-context Learning

Questions

- Which learning mode is more proficient in a language L?
- Do they differ in their inductive bias?

Three Desiderata

- Can we precisely define the language?
- Can we compare language proficiency?
- Do we allocate equal resources both modes?

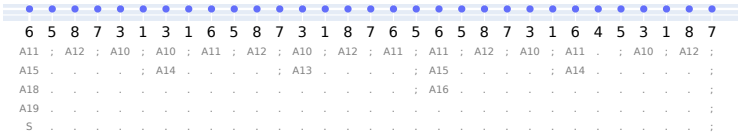
Three Desiderata

- Can we precisely define the language? → Formal language
- Can we compare language proficiency? → Discriminative test
- Do we allocate equal resources both modes? → Identical training/test sequences

Formal Language Learning

$S \rightarrow A19$ [1]
$A19 \rightarrow A18\ A16$ [0.50]
$A19 \rightarrow A16\ A18\ A17$ [0.50]
$A18 \rightarrow A15\ A14\ A13$ [0.50]
$A18 \rightarrow A14\ A15\ A13$ [0.50]
$A17 \rightarrow A14\ A13\ A15$ [0.50]
$A17 \rightarrow A13\ A14\ A15$ [0.50]
$A16 \rightarrow A14\ A15$ [0.50]
$A16 \rightarrow A15\ A14$ [0.50]
$A15 \rightarrow A11\ A12\ A10$ [0.50]
$A15 \rightarrow A12\ A10\ A11$ [0.50]
$A14 \rightarrow A11\ A10\ A12$ [0.50]
$A14 \rightarrow A10\ A11\ A12$ [0.50]
$A13 \rightarrow A10\ A12\ A11$ [0.50]
$A13 \rightarrow A12\ A11\ A10$ [0.50]
$A12 \rightarrow 9\ 8\ 7$ [0.50]
$A12 \rightarrow 8\ 7$ [0.50]
$A11 \rightarrow 6\ 5$ [0.50]
$A11 \rightarrow 6\ 4\ 5$ [0.50]
$A10 \rightarrow 3\ 1$ [0.50]
$A10 \rightarrow 1\ 2\ 3$ [0.50]

Probabilistic formal language L is a distribution of sequences s



$$P_L(s) = \frac{1}{2^{23}}$$

Formal Language Learning

$S \rightarrow A19$ [1]
$A19 \rightarrow A18\ A16$ [0.50]
$A19 \rightarrow A16\ A18\ A17$ [0.50]
$A18 \rightarrow A15\ A14\ A13$ [0.50]
$A18 \rightarrow A14\ A15\ A13$ [0.50]
$A17 \rightarrow A14\ A13\ A15$ [0.50]
$A17 \rightarrow A13\ A14\ A15$ [0.50]
$A16 \rightarrow A14\ A15$ [0.50]
$A16 \rightarrow A15\ A14$ [0.50]
$A15 \rightarrow A11\ A12\ A10$ [0.50]
$A15 \rightarrow A12\ A10\ A11$ [0.50]
$A14 \rightarrow A11\ A10\ A12$ [0.50]
$A14 \rightarrow A10\ A11\ A12$ [0.50]
$A13 \rightarrow A10\ A12\ A11$ [0.50]
$A13 \rightarrow A12\ A11\ A10$ [0.50]
$A12 \rightarrow 9\ 8\ 7$ [0.50]
$A12 \rightarrow 8\ 7$ [0.50]
$A11 \rightarrow 6\ 5$ [0.50]
$A11 \rightarrow 6\ 4\ 5$ [0.50]
$A10 \rightarrow 3\ 1$ [0.50]
$A10 \rightarrow 1\ 2\ 3$ [0.50]

Probabilistic formal language L is a distribution of sequences s

$$P_L(s) = \frac{1}{2^{23}}$$

Usefulness of formal languages

- Demarcation between sequences inside and outside the language
- Exact sampling of train and test sequences
- Elimination of semantic ambiguity and data contamination
- Control of language complexity

Language Proficiency Test

Given two LLMs M_1 and M_2, and a language L, which LLM is more proficient in L?

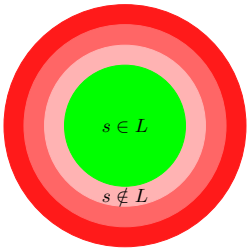
Language Proficiency Test

Given two LLMs M_1 and M_2, and a language L, which LLM is more proficient in L?

Can we directly compare generation probabilities of two LLMs?

Language Proficiency Test

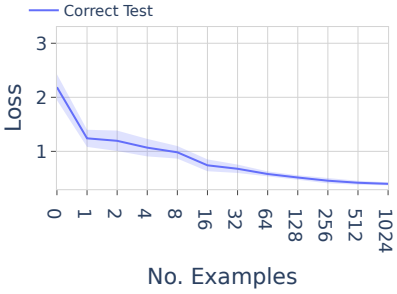
Given two LLMs M_1 and M_2, and a language L, which LLM is more proficient in L?



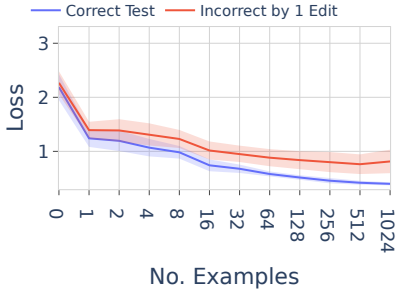
Can we directly compare generation probabilities of two LLMs?

Can we consider sequences outside L for language proficiency?

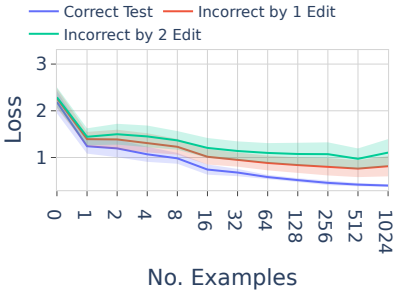
Evidence of Language Proficiency



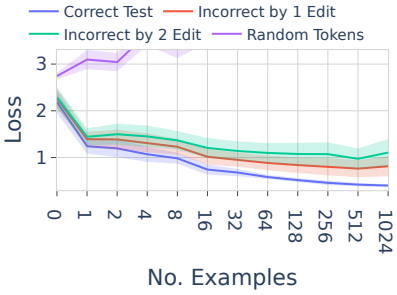
Evidence of Language Proficiency



Evidence of Language Proficiency



Evidence of Language Proficiency



Two Tests for Language Proficiency

Generative Test

With what probability does the LLM generate sequences in the language?

Compare generation probability across LLMs

Discriminative Test

Is generation probability higher for sequences in the language than sequences outside the language?

Compare discrimination success across LLMs

Two Tests for Language Proficiency

Generative Test

With what probability does the LLM generate sequences in the language?

Compare generation probability across LLMs

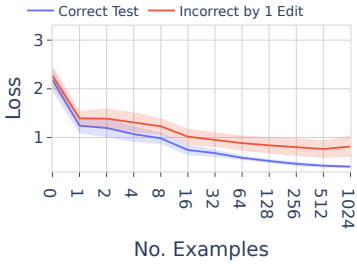
Discriminative Test

Is generation probability higher for sequences in the language than sequences outside the language?

Compare discrimination success across LLMs

The discriminative test avoids model-specific and prompt-specific biases

Generative vs. Discriminative Tests

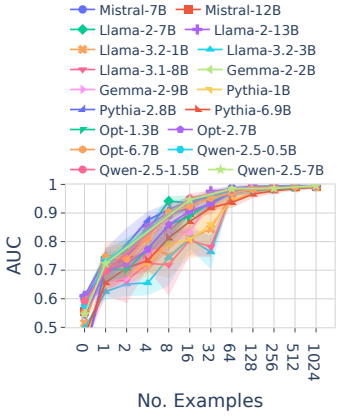


Generative Test

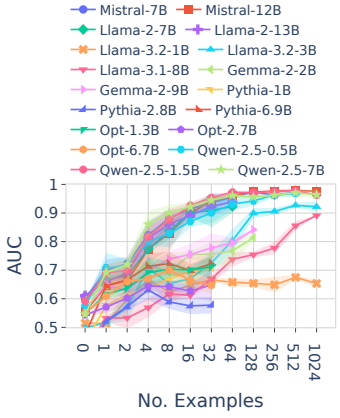
Discriminative Test

The hardest discriminative test: **correct sequence** vs. **sequence incorrect by edit distance** 1

Fine-tuning Converges across LLMs while In-context Learning Varies

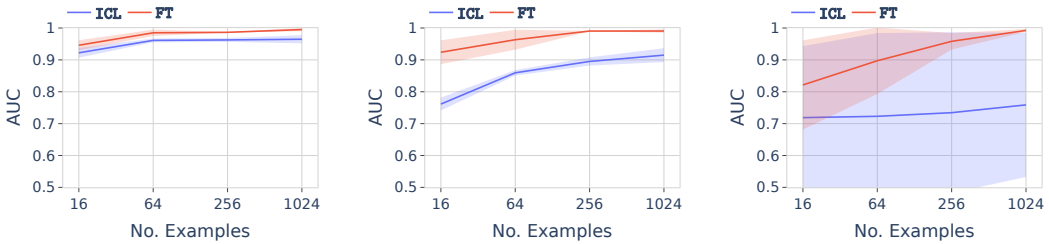


Fine-tuning



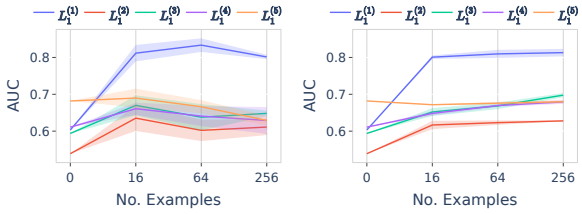
In-context Learning

Comparing Fine-tuning vs. In-context Learning: In-distribution



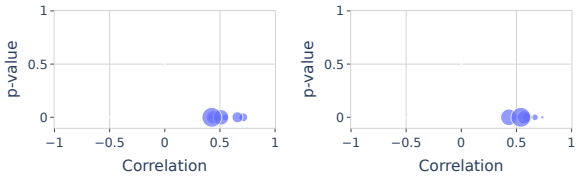
Fine-tuning is better than in-context learning on in-distribution language generalization

Comparing Fine-tuning vs. In-context Learning: Out-of-distribution



Both learning modes perform equally well and generalize to closer languages only

Inductive Bias of Learning Modes



Mistral-7B Llama-2-7B

Inductive bias is similar at low proficiency (fewer examples) but diverges at higher proficiency

Summary of Findings

- A **discriminative test** is comparable across LLMs; a generative test is not
- **Fine-tuning** $>$ **in-context learning** on in-distribution generalization
- **Both modes generalize equally** to closer out-of-distribution languages
- Inductive biases are **similar at low proficiency** but **diverge** at higher proficiency

Future Work

- Interplay of Learning & Memorization
- Identifying Implicit World Models
- A Formal Language Benchmark

Future Work

- Interplay of Learning & Memorization

When an LLM generates a training sequence,
can we disentangle memorization from learning?

- Identifying Implicit World Models

Which grammar does the LLM learn?

- A Formal Language Benchmark

Can we predict architectural choices of LLMs using formal language learning?

Future Work

- Interplay of Learning & Memorization

When an LLM generates a training sequence,
can we disentangle memorization from learning?

- Identifying Implicit World Models

Which grammar does the LLM learn?

- A Formal Language Benchmark

Can we predict architectural choices of LLMs using formal language learning?

Thank you!