Is In-Context Learning Sufficient for Instruction Following in LLMs?

Hao Zhao, Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion

EPFL

Background

1. Zhou et al. (2023) propose **Superficial Alignment Hypothesis**, which suggests that a few high-quality examples are sufficient to teach pre-trained LLMs to follow natural human instructions.
2. A line of work (Zhao et al., 2024) shows that IFT with 1K examples outperforms IFT with full datasets.
3. IFT of pre-trained LLMs permanently modifies model parameters, which causes huge costs for diverse use cases.
4. Lin et al. (2024) proposed **URIAL**, a method using *three* in-context examples to align base LLMs, achieving non-trivial instruction following performance.
5. ICL allows LLMs to learn from examples without changing model weights and offers flexible model preferences for different applications.
6. In particular, ICL is a promising capability for *long-context* LLMs that can potentially learn from *many* examples.

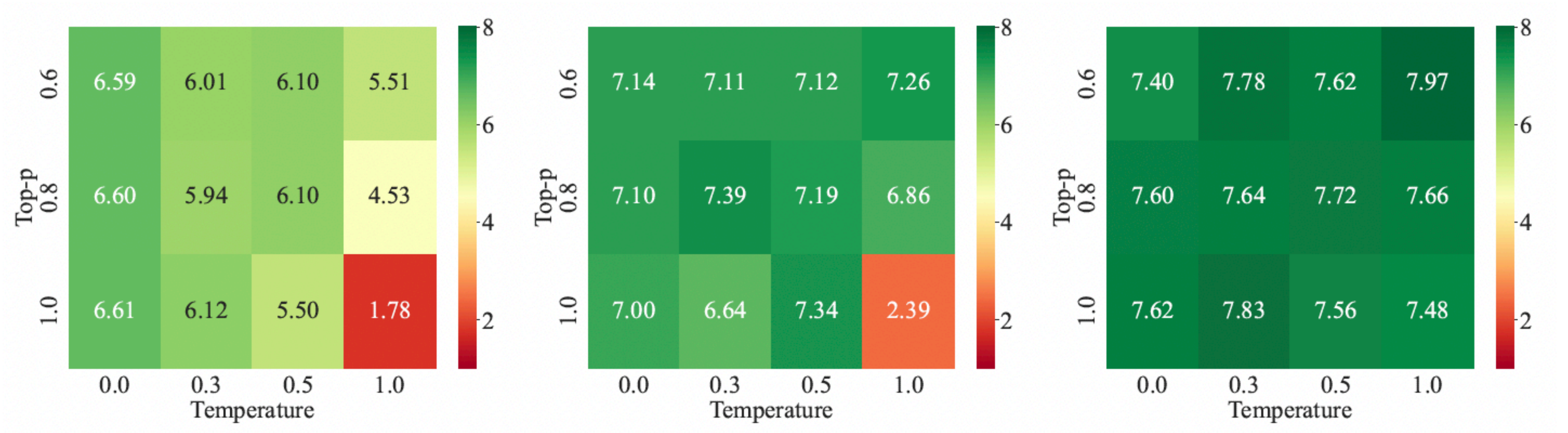
URIAL v.s. Aligned LLMs

- Firstly, we conduct systematic comparison of URIAL to aligned models on MT-Bench across different base LLMs, including GPT-4-Base.
- We show that URIAL still significantly lags behind aligned models fine-tuned with more sophisticated approaches.

Model	1st-turn	2nd-turn	Average
Llama-2-7B + URIAL *	5.75	3.91	4.83
Llama-2-7B-Instruct	**7.14**	**5.91**	**6.53**
Llama-2-70B + URIAL *	**7.61**	6.61	7.11
Llama-2-70B-Instruct	7.37	**7.03**	**7.20**
Llama-3-8B + URIAL *	6.84	4.65	5.75
Llama-3-8B-Instruct	**8.29**	**7.42**	**7.86**
Llama-3-70B + URIAL *	7.71	5.09	6.40
Llama-3-70B-Instruct	**8.96**	**8.51**	**8.74**
Llama-3.1-8B + URIAL *	6.95	5.31	6.13
Llama-3.1-8B-Instruct	**8.27**	**7.73**	**8.00**
Mistral-7B-v0.1 + URIAL *	**7.49**	5.86	6.67
Mistral-7B-Instruct-v0.1	7.31	**6.39**	**6.85**
Mistral-7B-v0.2 + URIAL *	6.99	5.55	6.27
Mistral-7B-Instruct-v0.2	**8.06**	**7.21**	**7.64**
Mixtral-8x22B-v0.1-4bit + URIAL	8.28	7.14	7.71
Mixtral-8x22B-Instruct-v0.1-4bit	**8.78**	**8.25**	**8.52**
GPT-4-Base + URIAL	7.96	5.04	6.50
GPT-4 (March 2023) *	**8.96**	**9.03**	**8.99**

Decoding Parameters

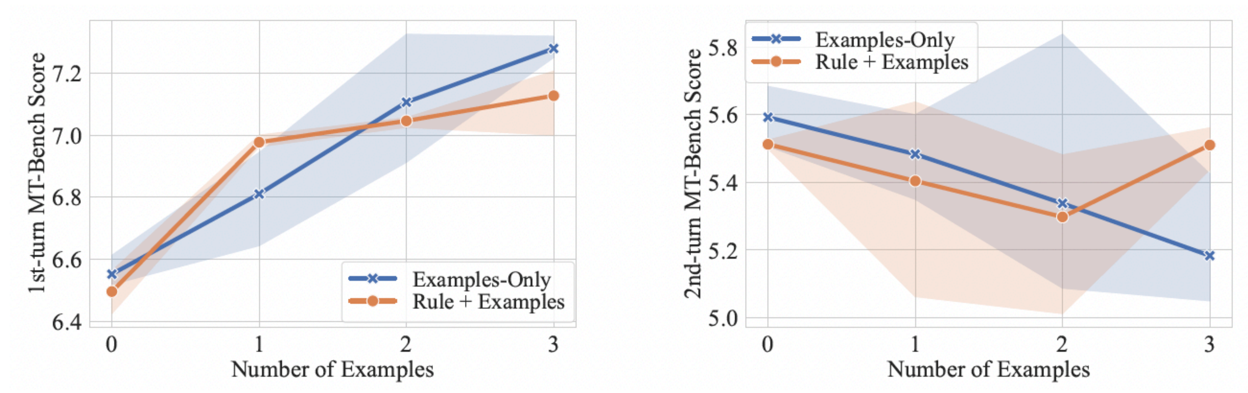
- We find that proper decoding schemes enable base LLMs to achieve reasonable instruction-following performance on MT-Bench.



(a) Base model (b) Base model with URIAL (c) Instruct model

A closer look at URIAL components

- Increasing the number of in-context examples progressively improves the performance of the base LLM.



Scaling up In-Context examples

Setups:

- Base models: Mistral-7B-v0.2 (32k) and Llama-3.1-8B (128k)

Strategies to select additional in-context examples:

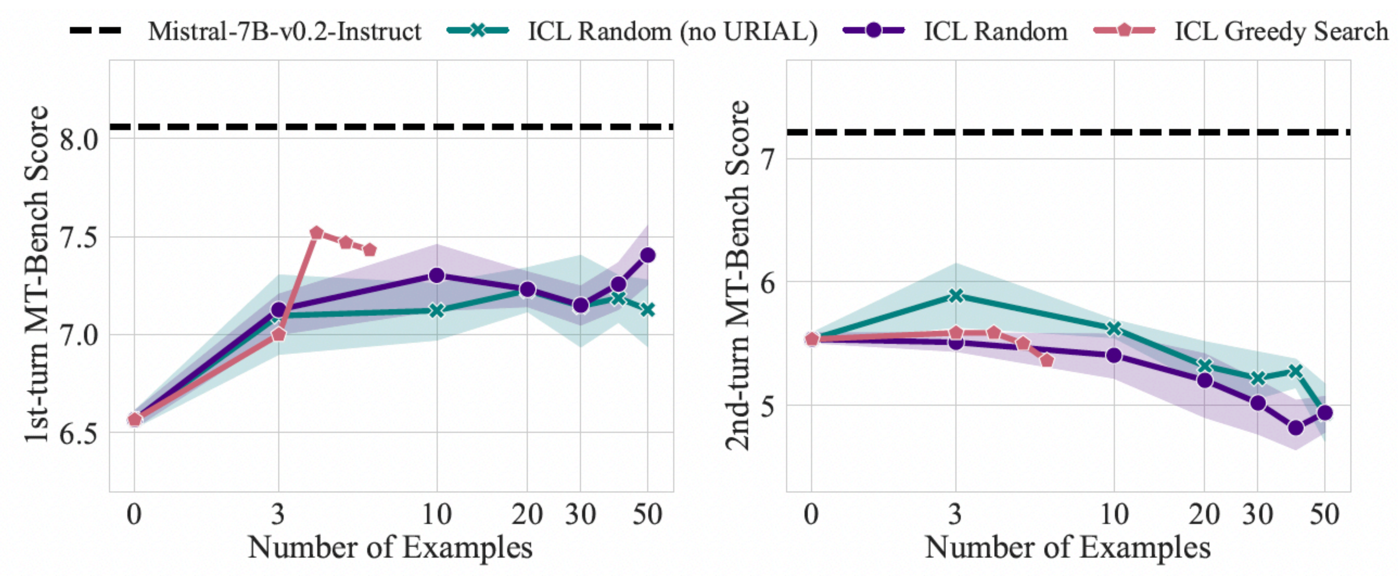
- Greedy search: select examples that greedily maximize the MT-Bench score using GPT-4-Turbo as the judge.
- Sampling from IFT datasets: Instruct-SkillMix contains high-quality examples.

Results:

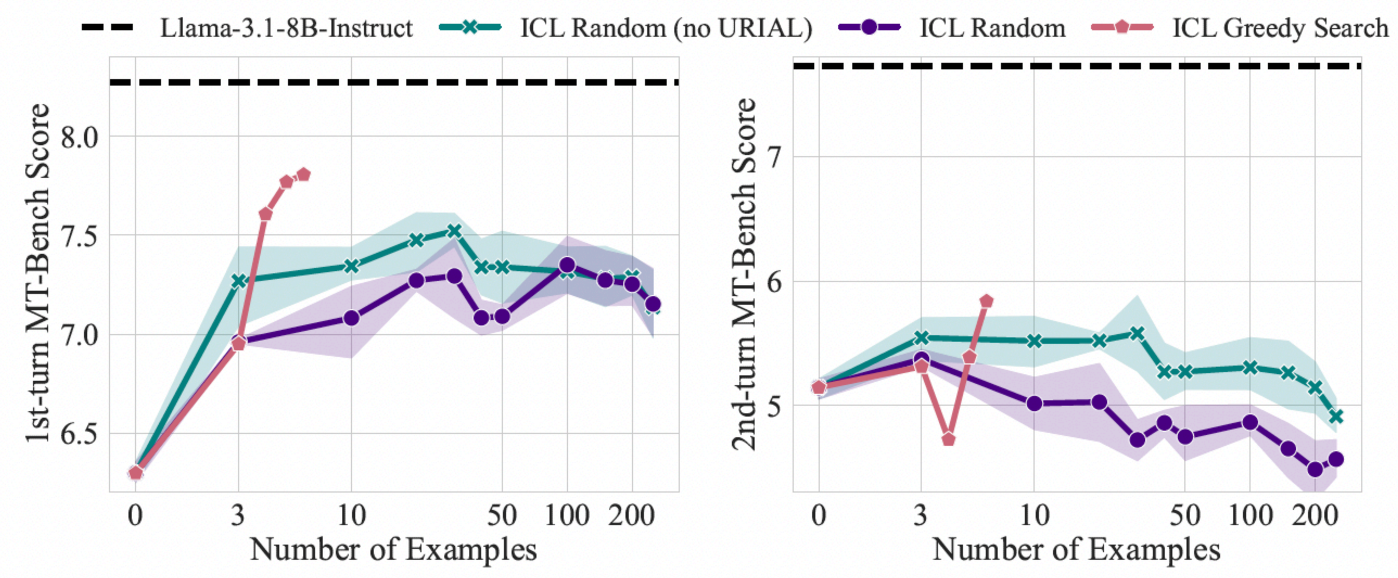
- Improved in-context alignment by greedy search

Model	Mistral-7B-v0.2		Llama-3.1-8B	
	MT-Bench (1st)	AlpacaEval 2.0	MT-Bench (1st)	AlpacaEval 2.0
URIAL (3 examples)	7.00	8.22	6.95	7.28
URIAL + greedy search (1 ex.)	**7.52**	7.53	7.61	**8.61**
URIAL + greedy search (2 ex.)	7.47	7.78	7.77	8.16
URIAL + greedy search (3 ex.)	7.43	**8.55**	**7.81**	8.19

- Scaling up in-context examples



(a) Mistral-7B-v0.2



(b) Llama-3.1-8B

Conclusions:

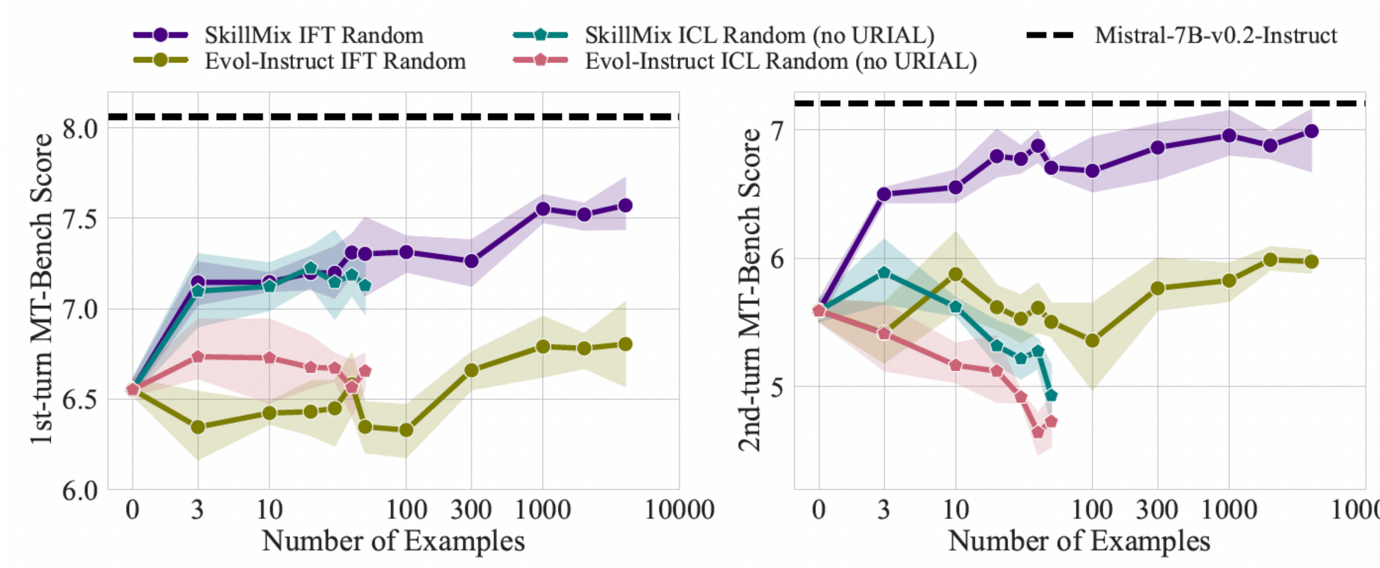
- Many-shot ICL can improve instruction following performance but fails to close the gap with aligned LLMs.
- The data selection scheme via greedy search outperforms, with 1 to 3 additional examples, the many-shot approach with random samples.

ICL vs IFT for Instruction Following

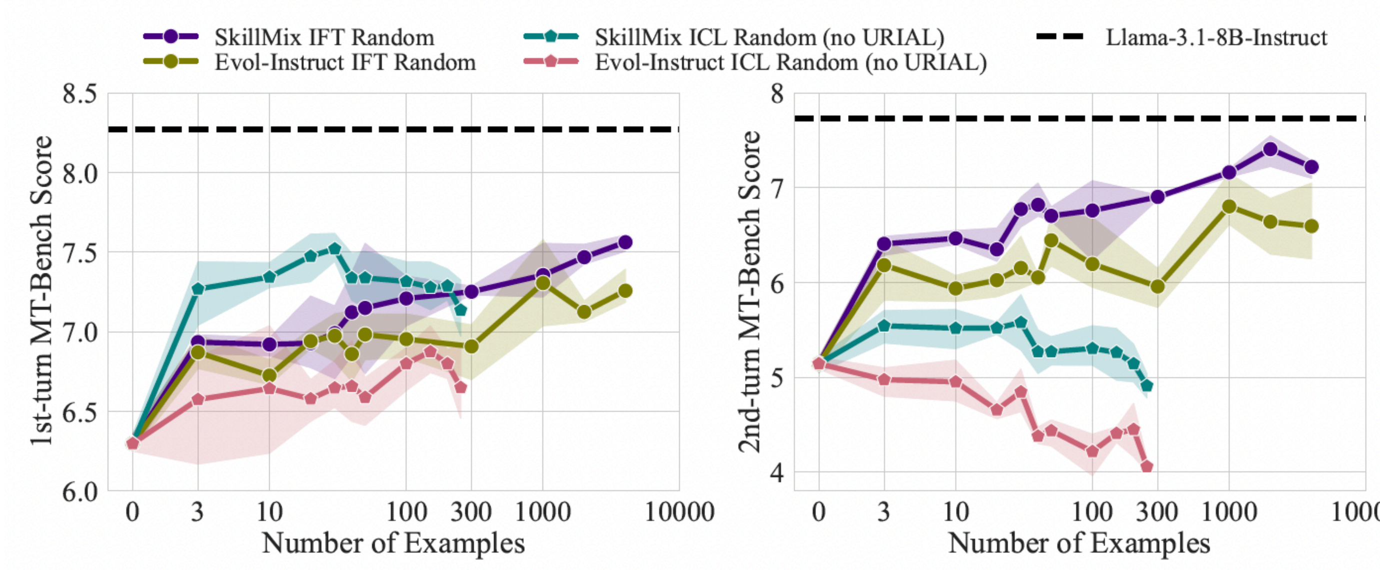
Setups:

- Base models: Mistral-7B-v0.2 (32k) and Llama-3.1-8B (128k)
- Datasets: Instruct-SkillMix (high quality), Evol-Instruct (medium quality)
- Fair comparison of ICL and IFT in the low-data regime, ranging from 3 to 4K.

Results:



(a) Mistral-7B-v0.2



(b) Llama-3.1-8B

Conclusions:

- We show that ICL and IFT with the same number of examples are roughly equivalent for single-turn conversations in the low-data regime.
- IFT generalizes substantially better than ICL when more examples are present, especially for multi-turn conversations.

1. Zhou, Chunting, et al. "Lima: Less is more for alignment." *Advances in Neural Information Processing Systems* 36 (2024).
2. Zhao, Hao, et al. "Long Is More for Alignment: A Simple but Tough-to-Beat Baseline for Instruction Fine-Tuning." *Forty-first International Conference on Machine Learning* (2024).
3. Lin, Bill Yuchen, et al. "The unlocking spell on base llms: Rethinking alignment via in-context learning." *The Twelfth International Conference on Learning Representations* (2023).