Small Language Models Improve Giants by Rewriting Their Outputs

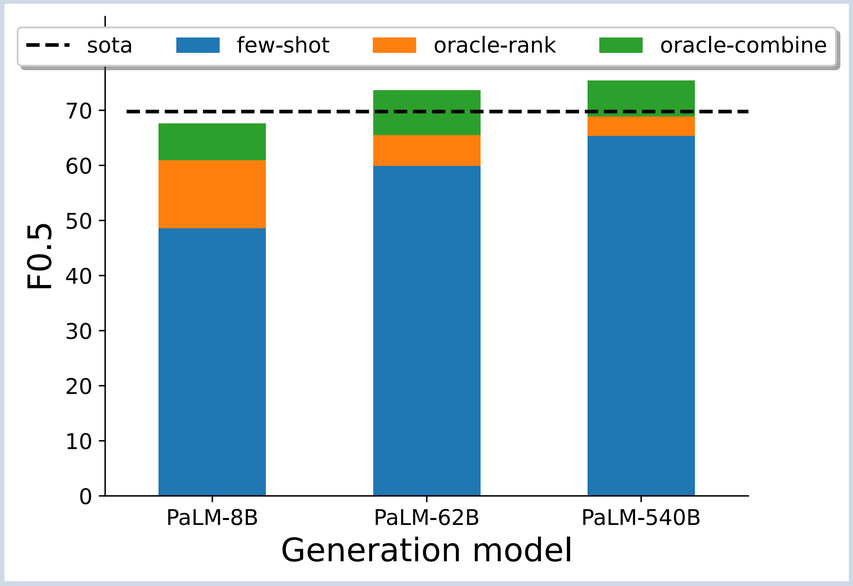
EACL 2024

Giorgos Vernikos, Arthur Bražinskas, Jakub Adamek, Jonathan Mallinson, Alieksei Severyn, Eric Malmi

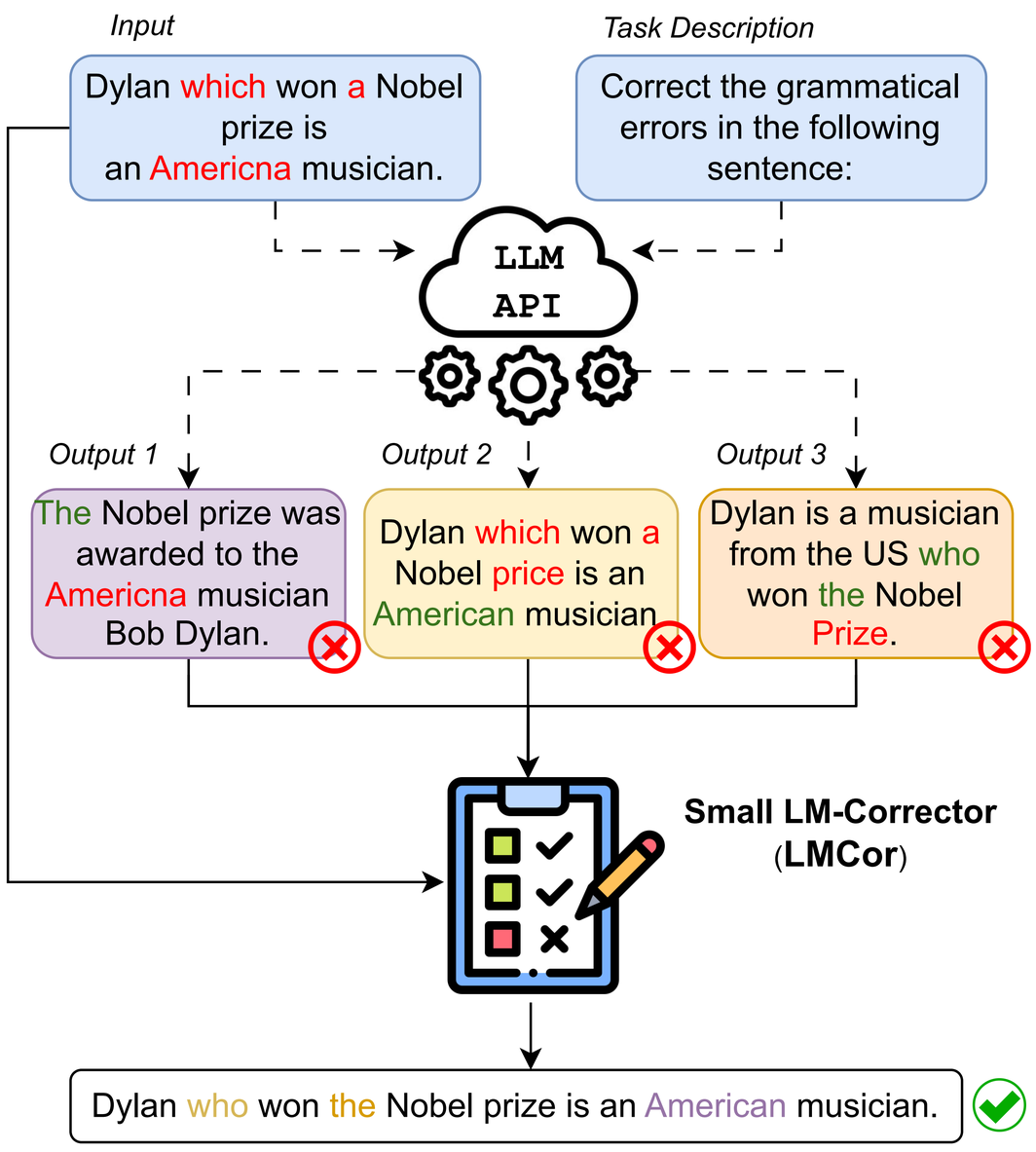
Research Question

Can we find a **parameter-efficient** way to improve the performance of LLMs **without access to their weights**?

Observation: LLMs can generate diverse candidates with *complementary strengths and weaknesses*.



LMCor

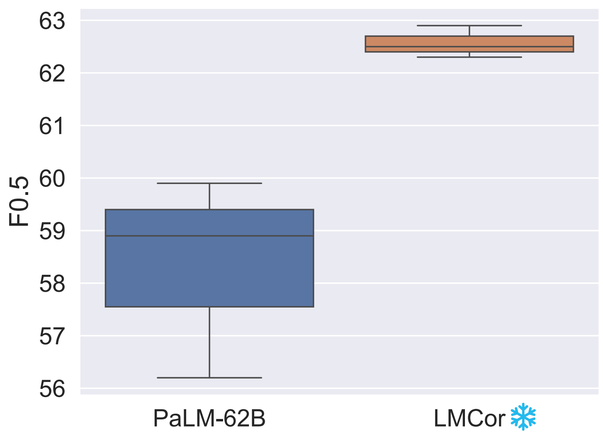


Experiments

Task	Size	LLM	LMCor	# cands
Grammatical error correction	60k	PaLM-62B	T5-base 250M	5
Data-to-Text generation	35k			
Summarisation	204k			
Machine translation	200k	XGLM-2.9B		

Analysis

Different in-context examples



Plug-and-play with various LLMs

same family, different scale

T5-base		59.38	
PaLM (ICL)	*8B*	*62B*	*540B*
	48.62	59.92	**65.37**
+ LMCor (single) ❄	61.40	**62.48**	63.55
+ LMCor (mult.) ❄	**61.89**	**62.47**	65.16

CoNLL-14

different family, different scale

Model	R-2	R-L
GPT3-Codex (ICL)*	34.2	44.4
+ MBRD-BLEURT*	36.4	46.5
+ LMCor (mult.) ❄	**44.8**	**53.0**

E2E NLG

❄ = frozen, no retraining

georgios.vernikos@epfl.ch @gvernikos https://georgevern.github.io/