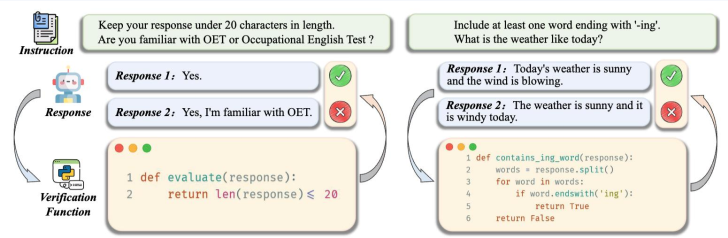
Self-play with Execution Feedback: Improving Instruction-following Capabilities of Large Language Models

Guanting Dong, Keming Lu, Chengpeng Li, Tingyu Xia, Bowen Yu†, Chang Zhou, Jingren Zhou

Qwen Team, Alibaba Inc.

ICLR 2025 (Spotlight, Top5% paper)

Motivation



A core strength of LLMs lies in executing natural language instructions. We presents AUTOIF , the first scalable framework for automated generation of high-quality instruction-following data. By reframing data validation as code verification, AUTOIF orchestrates three components: instruction generation, response-validation code synthesis, and unit test creation, forming a closed-loop quality assurance system. Execution feedback-driven rejection sampling efficiently produces data for SFT and RLHF. Evaluations on top open-source LLMs demonstrate substantial improvements across SFT, Offline/Online DPO training paradigms, particularly in self-alignment strong-to-weak distillation.

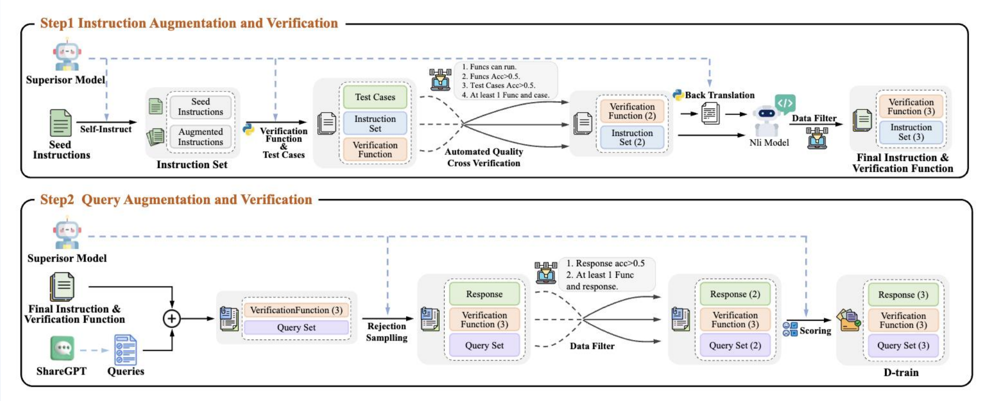
Contribution & Conclusion

➢ We propose AUTOIF to efficiently enhance LLMs' instruction - following. It converts instruction-following alignment into auto code verification, making LLMs generate instructions, verification code, and unit test samples.

➢ Based on DPO algorithms, we treat executor feedback as a reward model, create pairwise preference samples, and design offline/on-policy strategies to optimize the model's instruction-following.

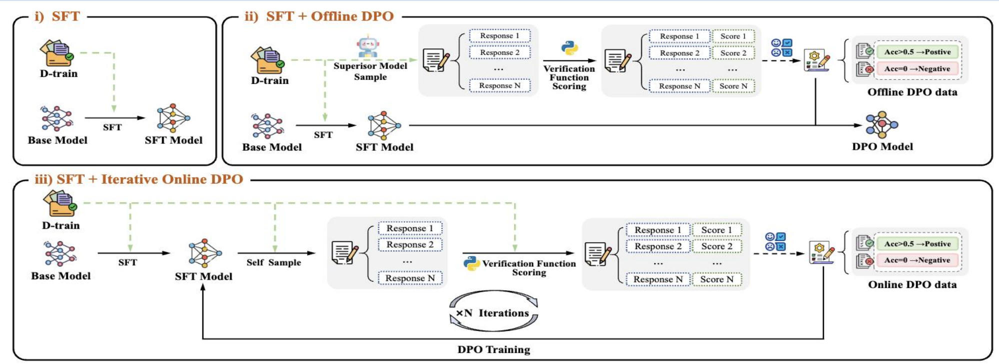
➢ AUTOIF is validated on benchmarks in "Self-Alignment" and "Strong-to-Weak" settings. It reaches over 90% accuracy in IFEval without sacrificing general and reasoning capability

Method: AUTOIF



1. Instruction Augmentation: Starting with seed instructions, LLMs generate augmented instructions through self-instruct.
2. Verification Functions: Automatically generating Python functions to verify the correctness of responses.
3. Back-Translation: Ensuring consistency between instructions and verification functions.
4. Query Augmentation: Creating diverse queries and responses for training.
5. Quality Filtering: Filtering data based on verification function accuracy and query relevance.

Training Strategies



Main Results

Model	IFEval				FollowBench (SSR)						C-Eval	MMLU	GSM8k	HumanEval
	Pr (S)	Pr. (L)	Ins. (S)	Ins. (L)	Level 1	Level 2	Level 3	Level 4	Level 5	Avg				
Baselines (< 10B)														
Qwen2-7B	37.7	43.6	49.4	53.4	55.6	53.5	53.7	49.9	48.6	52.3	74.4	64.4	71.1	58.1
Qwen2-7B(ShareGPT)	30.9	33.5	42.4	45.2	56.1	52.7	50.8	45.2	47.9	50.5	70.2	59.8	59.4	52.4
LLaMA3-8B	24.6	26.1	38.1	39.7	10.0	10.3	10.5	14.3	12.7	11.6	24.2	38.8	4.5	0.6
LLaMA3-8B(ShareGPT)	23.7	26.4	33.8	37.1	44.0	40.0	39.6	33.3	33.6	38.1	35.2	44.6	20.5	38.1
Mistral-7B	23.3	24.6	38.4	39.6	40.1	39.7	37.9	35.7	36.7	38.0	38.2	47.6	20.5	38.4
Baselines (> 10B)														
Qwen2-72B-Instruct	77.1	80.4	84.4	86.9	70.2	66.6	63.5	58.1	56.3	62.9	83.8	80.8	87.9	73.8
LLaMA3-70B-Instruct	77.8	83.8	84.2	88.8	60.7	60.5	61.1	61.7	60.3	60.9	60.2	80.5	92.6	78.7
Mixtral-8x22B	41.8	47.3	55.2	60.0	63.9	60.0	58.2	56.2	55.3	58.7	-	-	-	-
GPT-4†	76.9	79.3	83.6	85.4	84.7	77.6	76.2	77.9	73.3	77.9	-	-	-	-
GPT-3.5 Turbo†	-	-	-	-	80.3	71.2	74.2	69.6	67.1	72.5	-	-	-	-
Supervision Model: Qwen2-72B														
Strong-to-Weak														
Qwen2-7B-SFT	$40.7_{+3.0}$	$44.5_{+0.9}$	$51.3_{+1.9}$	$55.4_{+2.0}$	$60.2_{+4.6}$	$53.7_{+0.2}$	$54.3_{+0.6}$	$49.9_{+0.0}$	$48.6_{+0.0}$	$53.3_{+1.0}$	$73.9_{+0.0}$	$64.4_{+0.0}$	$74.1_{+3.0}$	$58.3_{+0.2}$
w/ Offline DPO	$41.2_{+3.5}$	$44.7_{+1.2}$	$51.4_{+2.0}$	$56.2_{+2.8}$	$61.4_{+5.8}$	$54.5_{+1.0}$	$54.3_{+0.6}$	$51.2_{+1.3}$	$48.6_{+0.0}$	$54.0_{+1.7}$	$75.1_{+0.7}$	$64.5_{+0.1}$	$72.9_{+1.8}$	$59.5_{+1.4}$
w/ Online DPO	$44.0_{+6.3}$	$46.6_{+3.0}$	$55.0_{+5.6}$	$57.9_{+4.5}$	$61.4_{+5.8}$	$56.8_{+3.3}$	$57.8_{+4.1}$	$55.4_{+5.5}$	$51.6_{+3.0}$	$56.6_{+4.3}$	$76.0_{+1.6}$	$64.8_{+0.4}$	$72.3_{+1.2}$	$58.2_{+0.1}$
Self-Alignment														
Qwen2-72B-Instruct w/ Online DPO	$80.2_{+3.1}$	$82.3_{+1.9}$	$86.1_{+1.7}$	$88.0_{+1.1}$	$76.2_{+6.0}$	$69.8_{+3.2}$	$67.0_{+3.5}$	$61.6_{+3.5}$	$62.8_{+6.5}$	$67.5_{+4.6}$	$84.9_{+1.1}$	$81.2_{+0.4}$	$88.2_{+0.3}$	$75.0_{+1.2}$
Supervision Model: LLaMA3-70B														
Strong-to-Weak														
LLaMA3-8B-SFT	$28.7_{+4.1}$	$40.3_{+14.2}$	$41.4_{+3.3}$	$52.2_{+12.05}$	$46.6_{+36.6}$	$46.2_{+35.9}$	$45.9_{+35.4}$	$37.6_{+23.3}$	$41.0_{+28.3}$	$43.5_{+31.9}$	$34.5_{+10.3}$	$45.6_{+6.8}$	$33.2_{+28.7}$	$38.2_{+37.6}$
w/ Offline DPO	$27.9_{+3.3}$	$41.6_{+15.5}$	$40.5_{+2.4}$	$54.1_{+14.4}$	$51.9_{+41.9}$	$51.3_{+41.0}$	$50.1_{+39.6}$	$45.3_{+31.0}$	$47.5_{+34.8}$	$49.2_{+37.6}$	$36.2_{+12.0}$	$45.3_{+6.5}$	$31.9_{+27.4}$	$38.5_{+37.9}$
w/ Online DPO	$28.8_{+4.2}$	$43.1_{+17.0}$	$42.2_{+4.1}$	$56.0_{+16.3}$	$54.6_{+44.6}$	$52.1_{+41.8}$	$50.0_{+39.5}$	$49.0_{+34.7}$	$43.7_{+31.0}$	$49.9_{+38.3}$	$38.2_{+14.0}$	$45.1_{+6.3}$	$32.5_{+28.0}$	$38.4_{+37.8}$
Self-Alignment														
LLaMA-3-70B w/ Online DPO	$80.2_{+2.4}$	$85.6_{+1.8}$	$86.7_{+2.5}$	$90.4_{+1.6}$	$71.0_{+10.3}$	$67.2_{+6.7}$	$66.2_{+5.1}$	$64.6_{+2.9}$	$63.5_{+3.2}$	$66.5_{+5.6}$	$61.6_{+1.4}$	$80.7_{+0.2}$	$92.7_{+0.1}$	$78.7_{+0.0}$

Results:
1. AUTOIF achieves up to 90.4% accuracy on IFEval and over 5% improvement on FollowBench.
2. No decline in other capabilities

Key Findings:
3. Online DPO outperforms Offline DPO by effectively targeting model weaknesses
4. Larger models (e.g., Qwen2-72B) show greater improvements.
5. Higher function pass rates lead to better performance.

Scaling Results