Dataset	Samples	Domain	Evaluation	Avg. Test Cases	Data Source	NL
JuICe (Agashe et al., 2019)	1,981	open	lexical	-	GitHub Notebooks	en
HumanEval (Chen et al., 2021)	164	4	execution	7.7	Hand-written	en
MBPP (Austin et al., 2021)	974	8	execution	3.0	Hand-written	en
APPS (Hendrycks et al., 2021)	10,000	0	execution	13.2	Competitions	en
DSP (Chandel et al., 2022)	1,119	16	execution	2.1	Github Notebooks	en
MTPB (Nijkamp et al., 2022)	115	8	execution	5.0	Hand-written	en
Exe-DS (Huang et al., 2022)	534	28	lexical	-	GitHub Notebooks	en
DS-1000 (Lai et al., 2022)	1,000	7	execution	1.6	StackOverflow	en
CoNaLa (Yin et al., 2018)	2,879	open	lexical	-	StackOverflow	en
MCoNaLa (Wang et al., 2022)	896	open	lexical	-	StackOverflow	es, ja, ru
ODEX	945	79	lexical execution	1.8	StackOverflow Hand-Written	en, es, ja, ru