Conference Paper Title*

1st Federico Raspanti
Eindhoven University of Technology
Université Côte d'Azur
City, Country
f.raspanti@student.tue.nl

***Abstract*—**
***Index Terms*—**

Dataset	FOLIOv2	PrOntoQA	ProofWriter
Logic-LM	59.25%—54.67%		
+Refinement	58.69%—58.57%		
DynoReasoner End2End	63.12%—62.93%		
+Refinement	62.16%		
+Constrained Generation	63.55%		
+Refinement & Constrained Generation			

TABLE I
ACCURACY OF EXECUTABLE SAMPLES (F1)

Dataset	**Logic-LM**	**+ Refinement**	**+ Dynamic Examples**	**+ Both**
FOLIOv2	61.57%	60.59%	**64.03%**	63.54%
PrOntoQA				
ProofWriter				

TABLE II
ACCURACY (F1) WITH FEW-SHOT COT BACKUP IF SAMPLES ARE NON-EXECUTABLE

Dataset	**Direct Few-shot**	**CoT Few-shot**
FOLIOv2	47.05%—42.85—41.37	66.17%—64.61%—66.12%—67.27%
PrOntoQA		
ProofWriter		

TABLE III
ACCURACY (F1) OF BACKUP ON NON-EXECUTABLE SAMPLES

Dataset	**Direct Few-shot**	**CoT Few-shot**
FOLIOv2	47.05%—42.85—41.37	66.17%—64.61%—66.12%—67.27%
PrOntoQA		
ProofWriter		

TABLE IV
ACCURACY (F1) OF BACKUP ON NON-EXECUTABLE SAMPLES

Dataset	**Logic-LM**	**+ Refinement**	**+ Dynamic Examples**	**+ Both**
FOLIOv2	66.50%—68.47%	67.98%—68.96%	69.45%—70.44%	**72.90%**
PrOntoQA				
ProofWriter				

TABLE V
FULLY EXECUTABLE SAMPLES RATE - SAMPLES THAT CAN BE BOTH PARSED AND EXECUTED (%)

Dataset	**Logic-LM**	**+ Refinement**	**+ Dynamic Examples**	**+ Both**
FOLIOv2	**13.79%**	13.79%—14.28%	15.27%—15.76%	14.77%
PrOntoQA				
ProofWriter				

TABLE VI
PARSING ERRORS RATE (%)

Dataset	**Logic-LM**	**+ Refinement**	**+ Dynamic Examples**	**+ Both**
FOLIOv2	19.70%	18.22%—16.74%	15.27%—12.80%	**12.31%**
PrOntoQA				
ProofWriter				

TABLE VII
EXECUTION ERRORS RATE (%)