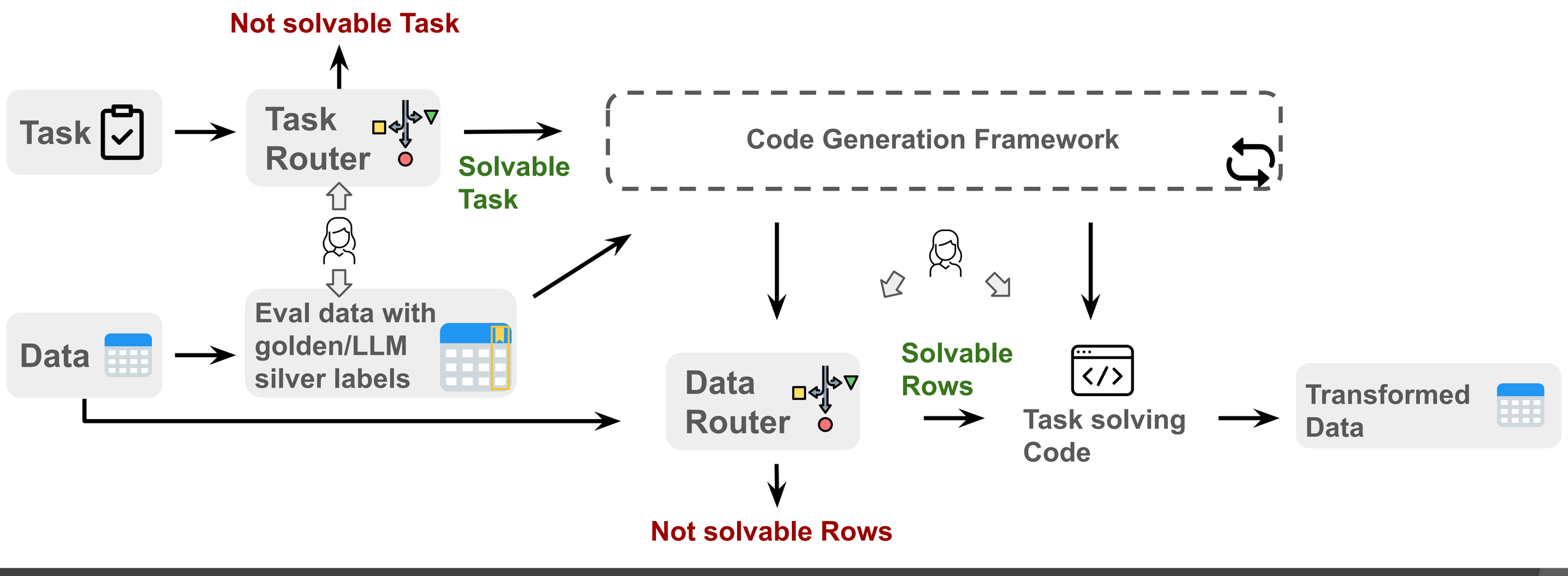
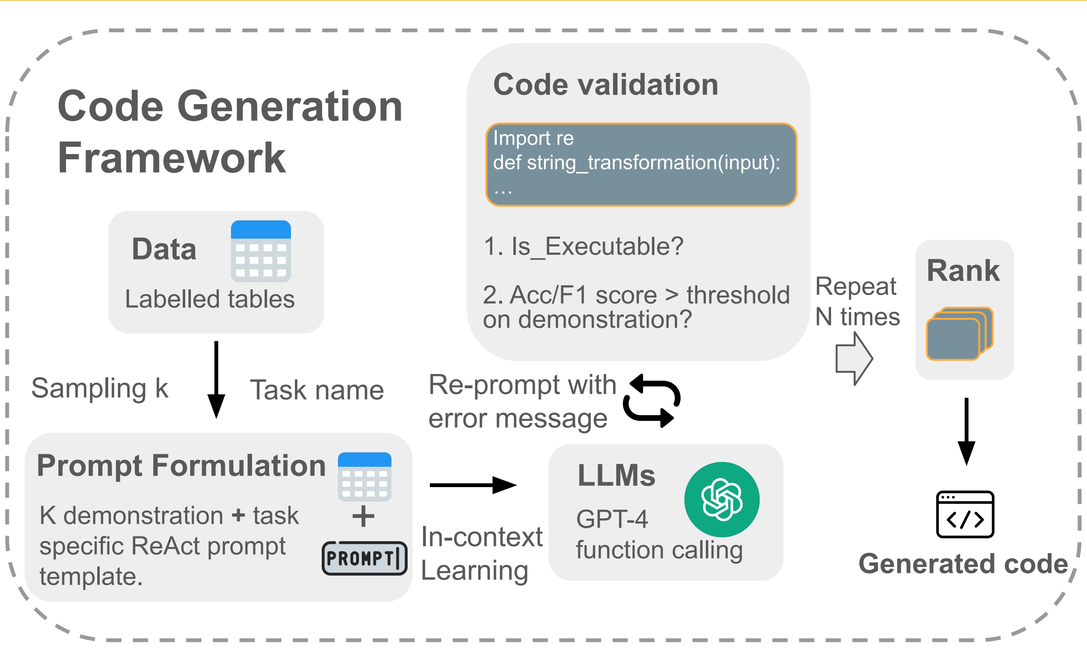
Towards Efficient Data Wrangling with LLMs using Code Generation

Xue Li, Till Döhmen



***Envisioned System Overview*: Task router directs task from solvable to not solvable; Data router routes data from solvable rows to not solvable. Code Generation Framework generates codes to wrangle your data.**



Code Generation Framework.

Task	Dataset	LLMPR[9]	Code Generation (Ours)	
EM	Fodors-Zagats	100	95.5	
EM	Beer	100	75.0	
EM	DBLP-ACM	96.6	19.7	76.9 ↓
EM	DBLP-GoogleScholar	83.8	69.7	
EM	Amazon-Google	63.5	42.1	
EM	iTunes-Amazon	98.2	70.0	
EM	Walmart-Amazon	87.0	25.5	
DI	Buy	98.5	84.6	
DI	Restaurant	88.4	50	
ED	Hospital	97.8	23.5	
ED	Adult	99.1	100*	

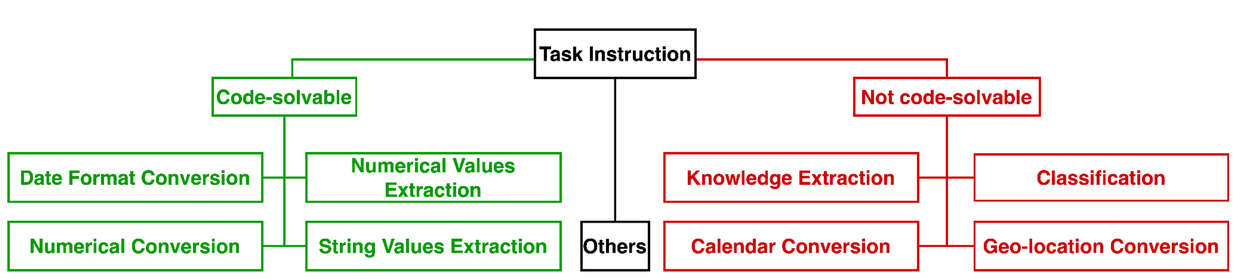
Results on Entity Matching, Data Imputation, Error Detection with Python.

Dataset	PBE [4]	LLMPR [9]	Code Generation (Ours)	
BingQL-semantics	32.0	54.0	**91.6**	37.6 ↑
BingQL-Unit	**96.0**	N/A	95.0	
Stack-overflow	63.0	65.3	**87.4**	
FF-GR-Trifacta	**91.0**	N/A	83.7	
Head cases	**82.0**	N/A	74.6	
Average	72.8	N/A	**86.46**	

Results on Data Transformation with Python.

	Bing-QL-semantics	Bing-QL-unit
GPT-4 (DuckDB-SQL)	65.3	96.0
GPT-4o (DuckDB-SQL)	67.3	**96.0**
GPT-4 (Python)	**91.6**	95.0

Results on Data Transformation with DuckDB SQL Macro.



Task Breakdown on 100 Data Transformation tasks.

MotherDuck

UNIVERSITY OF AMSTERDAM

(Effy) Xue Li
Corresponding Author
x.li3@uva.nl

Scan the QR code to check out our **github repo.**