LyS@SemEval-2025 Task 8: Zero-Shot Code Generation for Tabular QA

Adrian Gude, Roi Santos-Ríos, Francisco Prado-Valiño, Ana Ezquerro AND Jesús Vilares

Centro de Investigación en TIC (CITIC), Universidade da Coruña, Spain

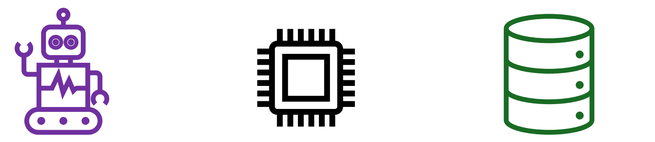
Tabular QA as Code Generation

Leverage powerful LLMs to *generate* executable code.

- Column selection to identify relevant variables and data types (Herzig et al., 2020).
- Code fixer module: capture and integrate runtime errors in the prompt.

Result: Guided prompt-based code generation (no training/finetuning).

Key components of the system

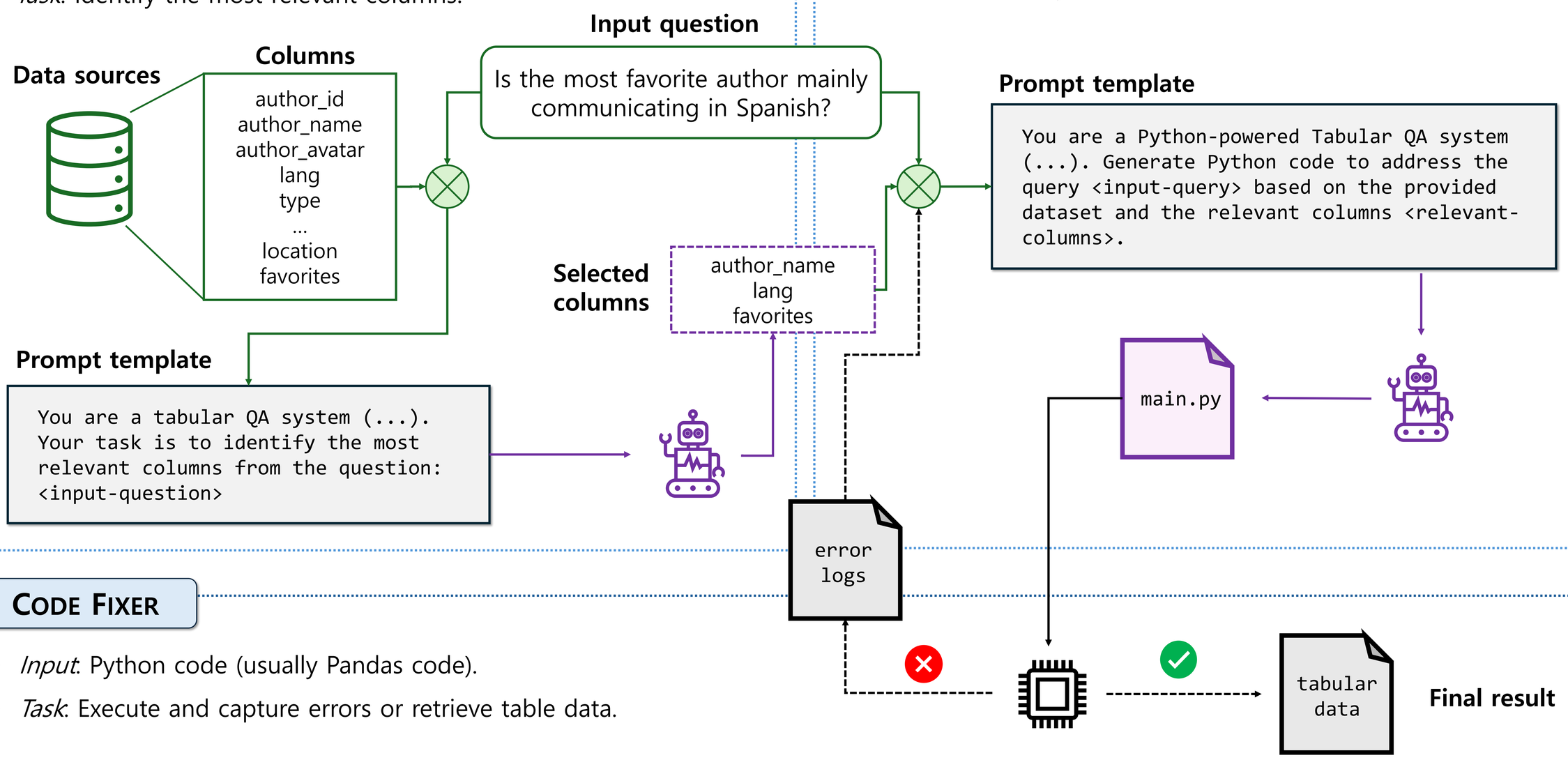


Column Selector

- *Input*: question (natural language) + list of columns.
- *Task*: Identify the most relevant columns.

Answer Generator

- *Input*: question (natural language) + selected columns.
- *Task*: Generate Python code.



Code Fixer

- *Input*: Python code (usually Pandas code).
- *Task*: Execute and capture errors or retrieve table data.

Experiments and Results

- Better performance than the baseline β (**S1**: 27.00; **S2**: 26.00).
- One of the best systems in the development phase (top 5).
- Ablation study with different LLMs.
  - Qwen-2.5-Coder (7B, 32B), Codestral (22B), Mistral (7B).
- Ablation study with different configurations:
  - Enhanced Column Selection (ECS)
  - Answer Generator (AG).
  - Code Fixer (CF).
- **Final phase**:
  - General category: 32nd (S1) and 31st (S2) out of 49 participants.
  - Open models category: 23rd (S1) and 21st (S2) out of 35 participants.

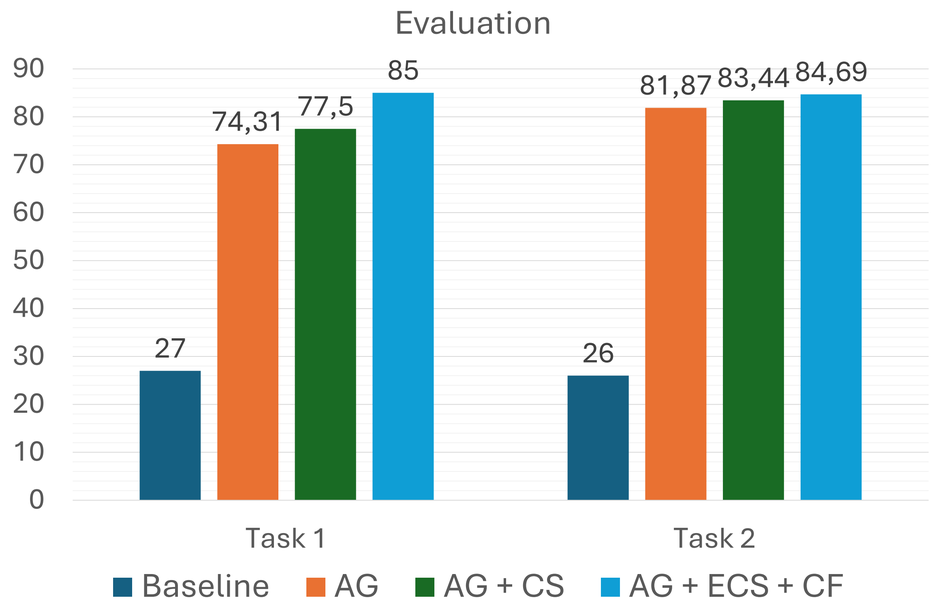
Ablation study with Qwen-2.5-Coder^{32B} on the dev set

		boolean	category	number	list[category]	list[number]	μ
S1	AG	81.25	78.12	75.00	65.62	70.31	74.06
	AG+CS	82.81	78.12	78.12	68.75	79.69	77.50
	AG+ECS+CF	**89.06**	**85.94**	**85.94**	**78.12**	**85.94**	**85.00**
S2	AG	84.37	**89.06**	85.94	75.00	75.00	81.87
	AG+CS	84.37	**89.06**	**90.62**	73.44	**79.69**	83.44
	AG+ECS+CF	**89.06**	**89.06**	**90.62**	**76.56**	78.12	**84.69**

Ablation study of different LLMs on the dev set (only Answer Generator)

		boolean	category	number	list[category]	list[number]	μ	β
S1	Qwen-2.5-Coder^{7B}	67.19	68.75	75.00	3.12	3.12	43.44	27.00
	Mistral^{7B}	51.56	59.37	73.44	35.94	34.37	50.94	
	Codestral^{22B}	73.44	**82.81**	**82.81**	48.44	48.44	67.19	
	Qwen-2.5-Coder^{32B}	**81.25**	78.12	75.00	**65.62**	**70.31**	**74.06**	
S2	Qwen-2.5-Coder^{7B}	81.25	84.37	85.93	6.25	1.56	51.87	26.00
	Mistral^{7B}	46.87	56.25	65.62	32.81	25.00	45.31	
	Codestral^{22B}	71.87	**89.06**	84.37	53.12	60.94	71.87	
	Qwen-2.5-Coder^{32B}	**84.37**	**89.06**	**85.94**	**75.00**	**75.00**	**81.87**	

Average performance (μ) and baseline (β) included. Best performance in bold.



We acknowledge grants SCANNER-UDC (PID2020-113230RB-C21) funded by MICIU/AEI/10.13039/501100011033; GAP (PID2022-139308OA-I00) funded by MICIU/AEI/10.13039/501100011033/ and ERDF, EU; LATCHING (PID2023-147129OB-C21) funded by MICIU/AEI/10.13039/501100011033 and ERDF, EU; CIDMEFEO funded by the Spanish National Statistics Institute (INE); as well as funding by Xunta de Galicia (ED431C 2024/02), and Centro de Investigación de Galicia "CITIC", funded by the Xunta de Galicia through the collaboration agreement between the Consellería de Cultura, Educación, Formación Profesional e Universidades and the Galician universities for the reinforcement of the research centres of the Galician University System (CIGUS). CITIC, as a center accredited for excellence within the Galician University System and a member of the CIGUS Network, receives subsidies from the Department of Education, Science, Universities, and Vocational Training of the Xunta de Galicia. Additionally, it is co-financed by the EU through the FEDER Galicia 2021-27 operational program (Ref.ED431G 2023/01).