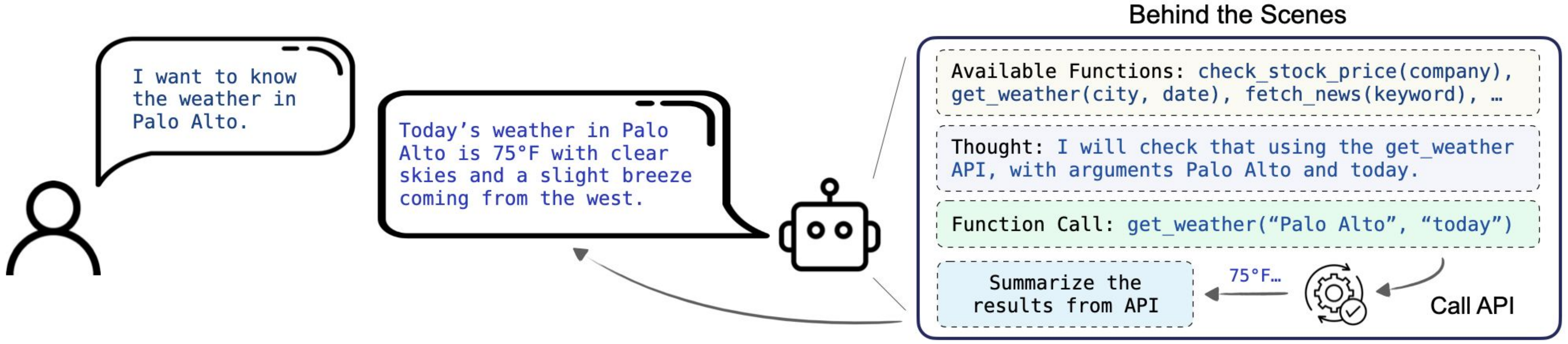
<u>APIGen</u>: **<u>A</u>**utomated **<u>PI</u>**peline for **<u>Gen</u>**erating Verifiable and Diverse Function-Calling Datasets

Zuxin Liu, Thai Hoang, Jianguo Zhang, Ming Zhu, Tian Lan, Shirley Kokane, Juntao Tan, Weiran Yao, Zhiwei Liu, Yihao Feng, Rithesh Murthy, Liangwei Yang, Silvio Savarese, Juan Carlos Niebles, Huan Wang, Shelby Heinecke, Caiming Xiong

<u>Salesforce AI Research, USA</u>

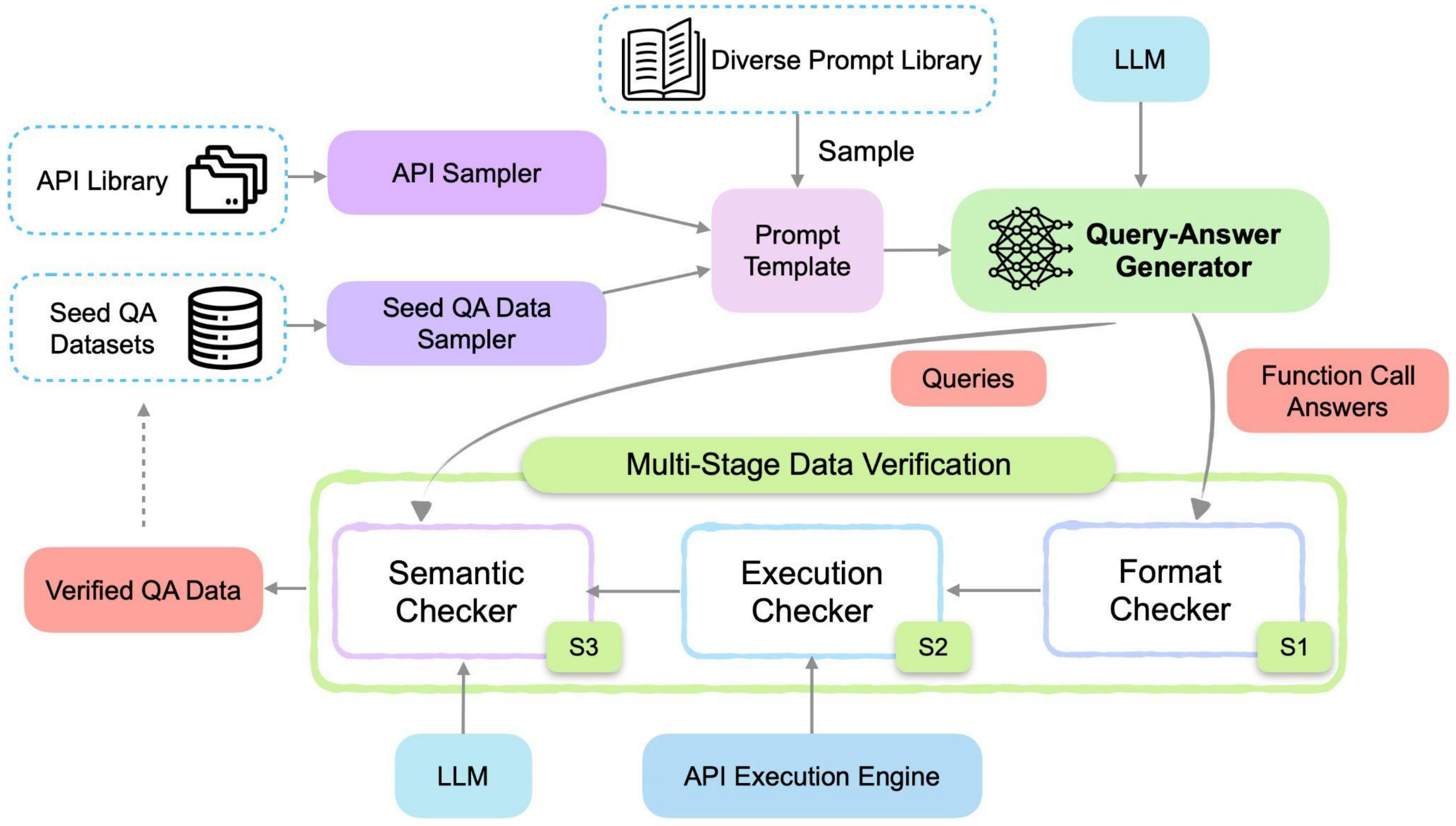
1. Introduction

Function-calling agents enable large language models (LLMs) to execute API calls based on natural language instructions. However, the effectiveness of these agents is often limited by the quality of training datasets, which tend to be static and lack verification.


Behind the Scenes
I want to know the weather in Palo Alto.
Today's weather in Palo Alto is 75°F with clear skies and a slight breeze coming from the west.
Available Functions: check_stock_price(company), get_weather(city, date), fetch_news(keyword), …
Thought: I will check that using the get_weather API, with arguments Palo Alto and today.
Function Call: get_weather("Palo Alto", "today")
Summarize the results from API
75°F…
Call API

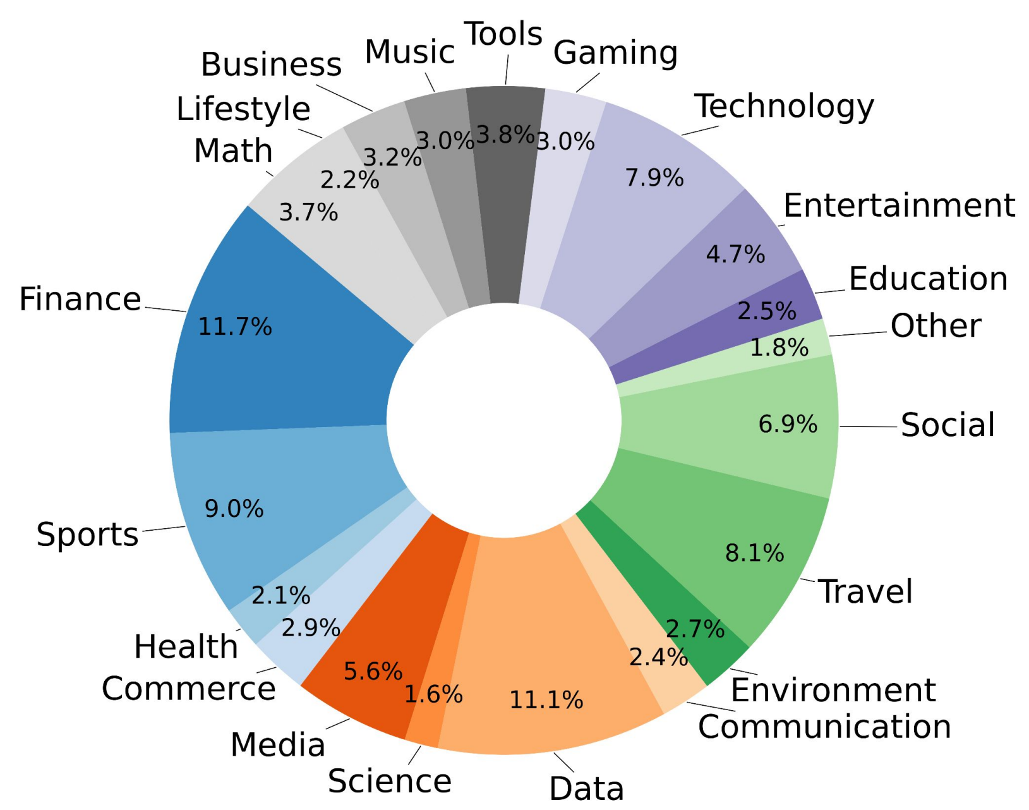
- We introduce **<u>APIGen</u>**, an **<u>A</u>**utomated **<u>PI</u>**peline for **<u>Gen</u>**erating diverse, reliable, high-quality datasets for training function-calling agents
- We generate a dataset of 60,000 high-quality data points across 21 categories using APIGen. Models trained with this dataset achieve SOTA performance on the Berkeley Function-Calling Benchmark.
- We release the dataset to benefit the research community and facilitate future advancements in this field.

2. APIGen Framework



- APIGen is designed with three key factors: **data quality**, **diversity**, and **collection scalability**.
- It achieves these through **multi-stage verification, sampling diversity**, and **modular design.**

3. Dataset Collection



3673 executable APIs across 21 different categories

API Sources

Python Functions

REST APIs

API quality filtering

Accessibility testing

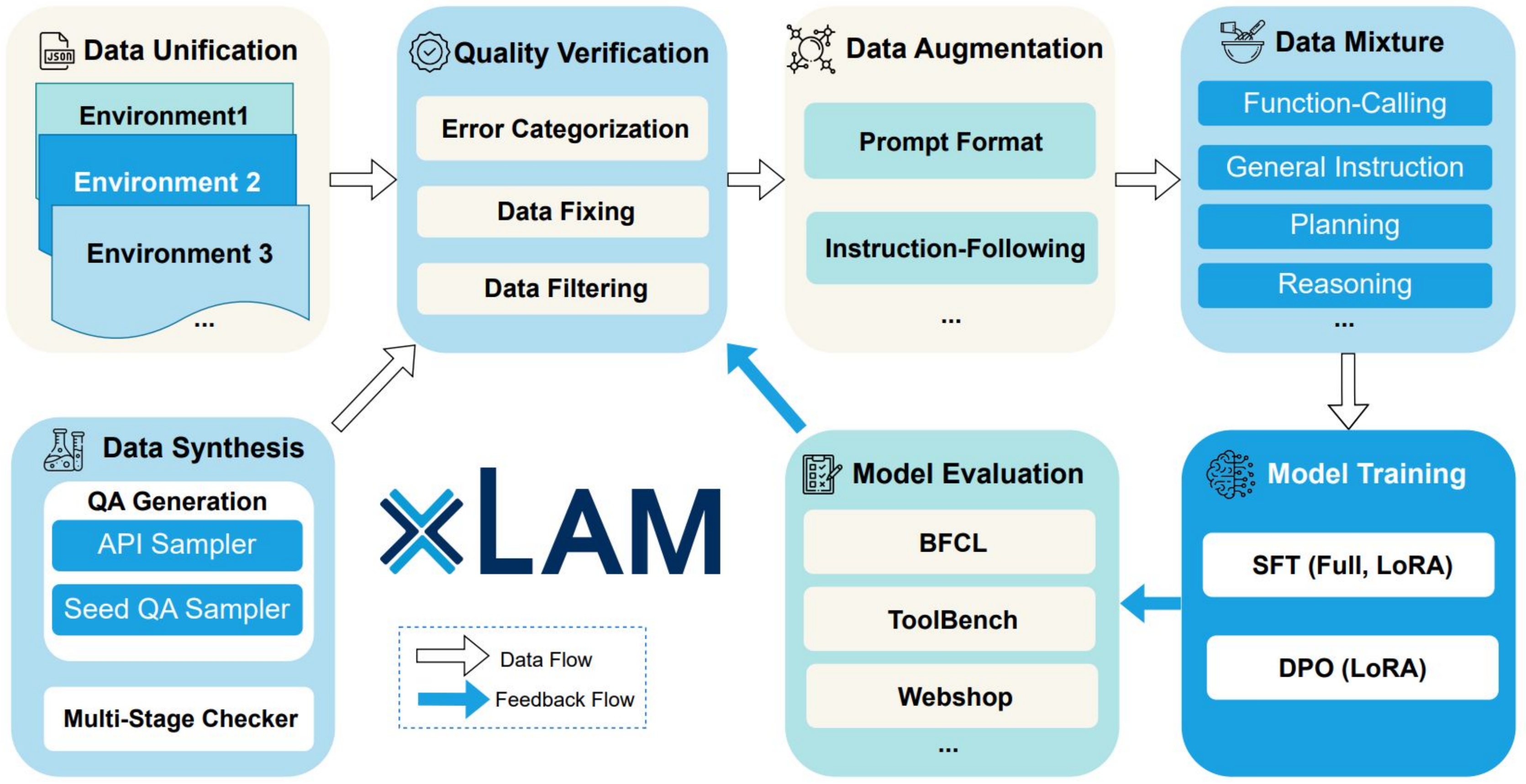
Docstring rewriting

...

Model	Verified Data	Fail Format	Fail Execution	Fail Semantic	Pass Rate
DeepSeek-Coder-33B-Inst	13,769	4,311	15,496	6,424	34.42%
Mixtral-8x7B-Inst	15,385	3,311	12,341	7,963	38.46%
Mixtral-8x22B-Inst	26,384	1,680	5,073	6,863	65.96%
DeepSeek-V2-Chat (236B)	33,659	817	3,359	2,165	84.15%

Filtering statistics for the generated datasets using different base LLMs. Stronger models demonstrated superior format-following capabilities and higher pass rates, suggesting strict verification is crucial for weaker models..

4. xLAM Model Training and Experiments



APIs

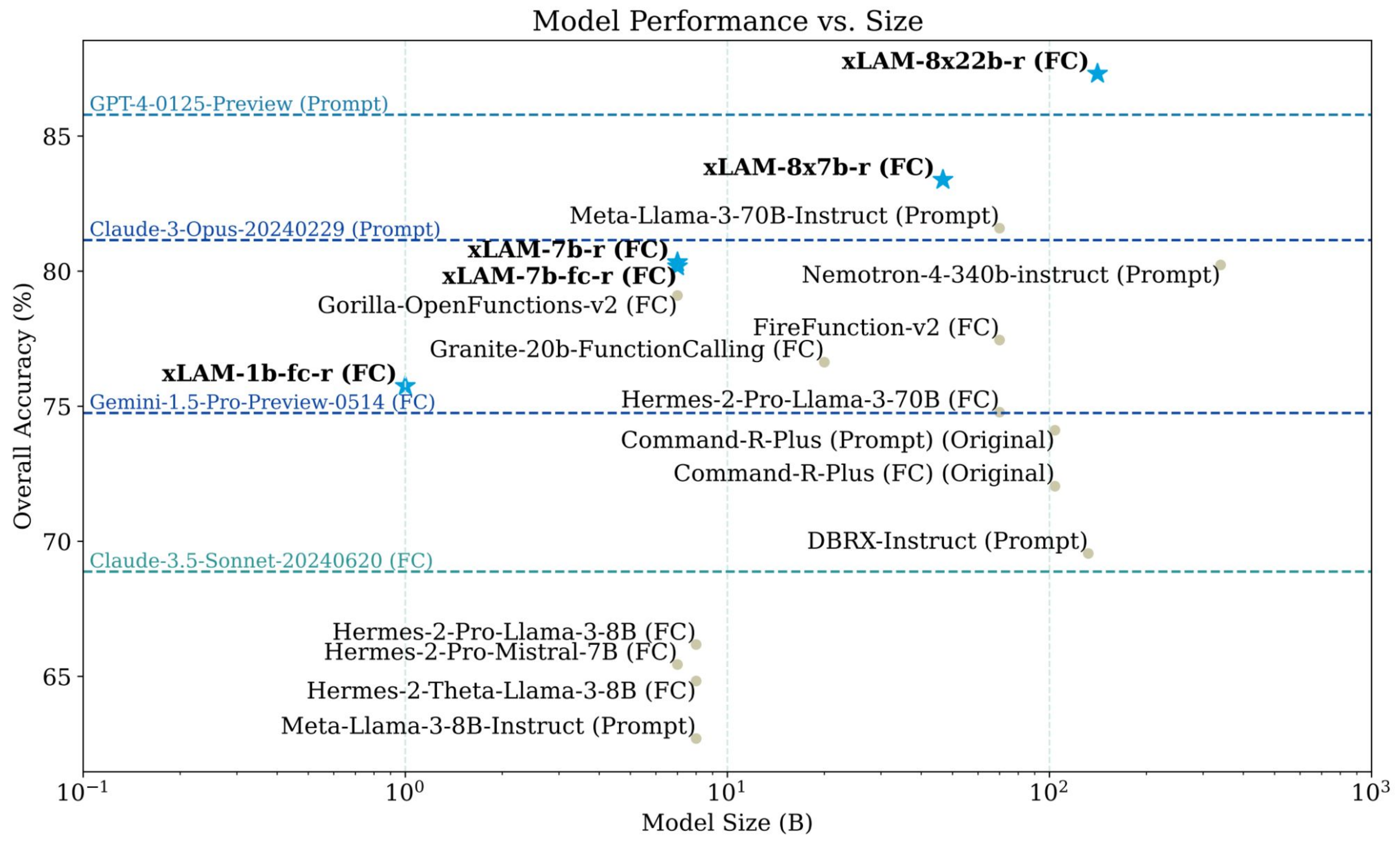
```
{
  "name": "api_name",
  "description": "API description",
  "parameters": {
    "param_name": {
      "type": "data type",
      "description": "",
      "default": "",
      "required": true or false
    },
    ... (more parameters)
  }
}
```

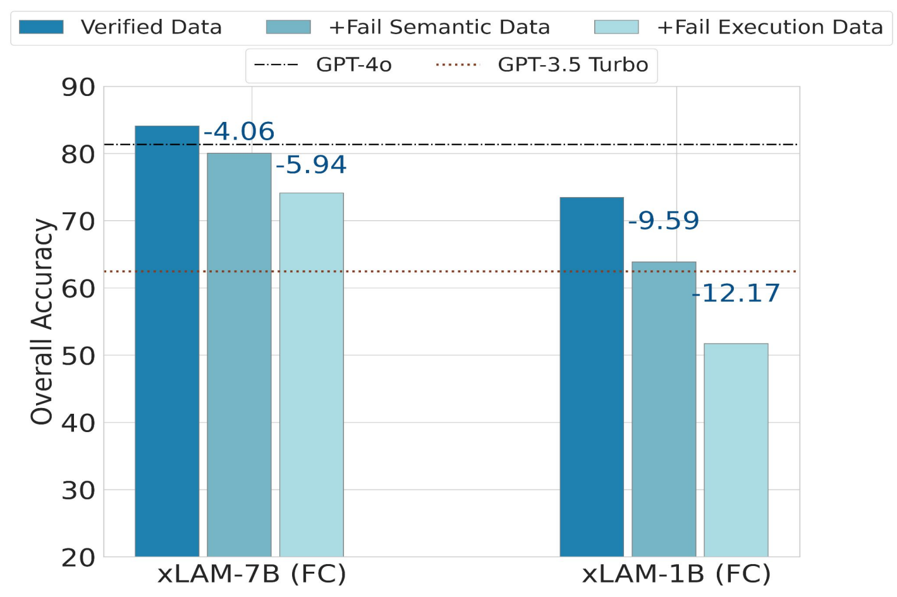
Function Call

```
[
 {
   "name": "api_name",
   "arguments": {
     "arg_name1": “value1",
     "arg_name2": "value2",
     ... (more arguments)
   }
 }
 ... (more API calls)
]
```

Generator Output

```
{
 "query": "The generated query.",
 "answers": [
   {
     "name": "api_name",
     "arguments": {
       "arg_name": "value",
       ... (more arguments)
     }
   }
   ... (more API calls)
 ]
}
```





- SOTA performance on Berkeley Function Calling Leaderboard (BFCL) v2
- Ablation studies show the importance of the verification process and data quality