MONAQ

Multi-Objective Neural Architecture Querying for Time-Series Analysis on Resource-Constrained Devices

Patara Trirat[1] Jae-Gil Lee[2]

[1] DeepAuto.ai [2] KAIST

30th ANNIVERSARY EMNLP 2025
Suzhou, China | 中国苏州
November 4-9 | 11月4日–9日

Motivation

Explosion of time-series data from IoT, wearables, and mobile devices.

Challenges

- Deploying DL models on MCUs (≤ 512kB SRAM) is non-trivial.
- Existing hardware-aware NAS (HW-NAS) methods:
  - Focus on vision tasks.
  - Require fixed search spaces and manual tuning.

Goal

Democratize efficient on-device time-series model design.

Research Questions

RQ1.
How to find high-performing architectures without user-defined search spaces?

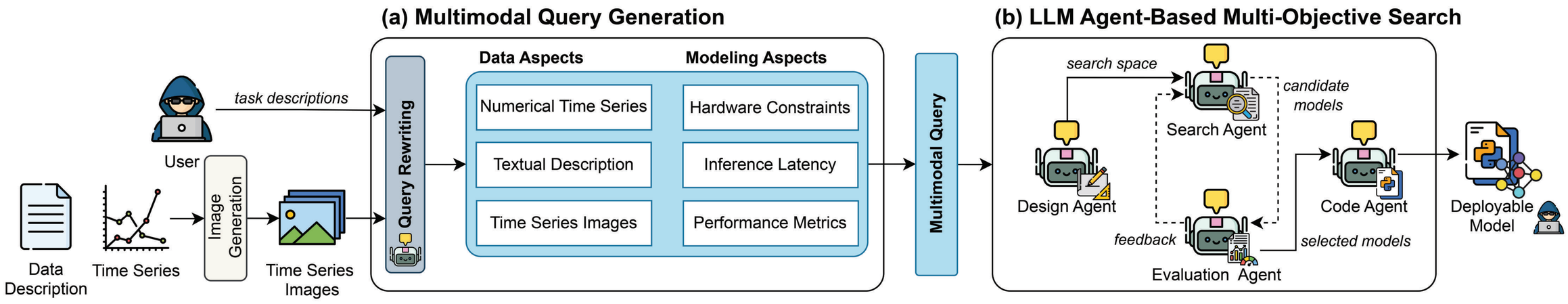
RQ2.
How to make LLM agents accurately understand time-series data and user requirements?

Our Contribution

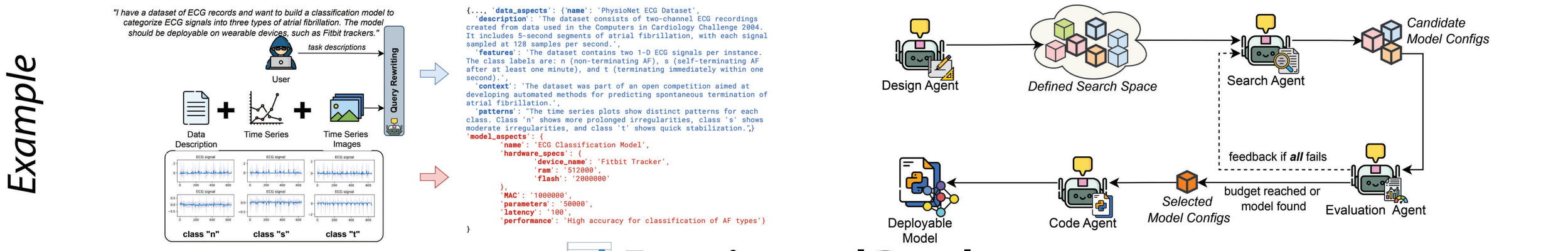
A LLM-driven framework that reformulates NAS as **Neural Architecture Querying (NAQ)** with:

- **Natural language input only**
  > no need for initial architectures or search spaces.
- **Multimodal query generation**
  > (numerical, textual, visual) for time-series.
- **Multi-agent LLMs**
  > to conduct low-cost, training-free multi-objective search.

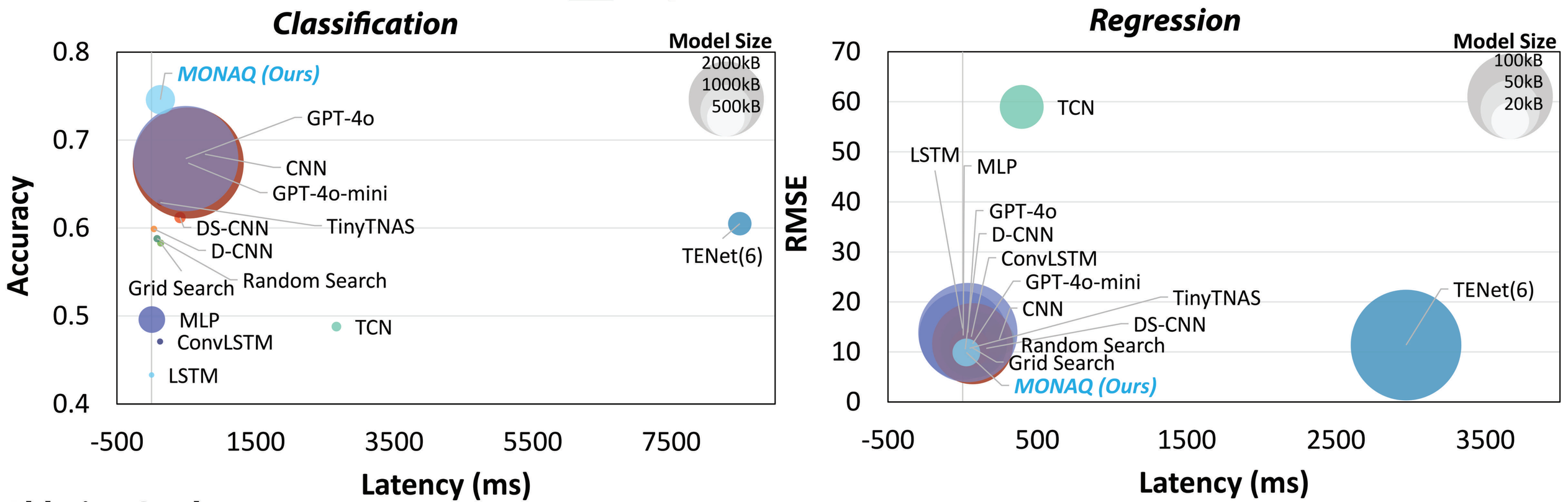
Framework Overview



Example



Experimental Results



Ablation Study

Table 2. Ablation study results on query rewriting and various agent combinations.

Variations	Classification			Regression		
	Latency (ms)	Accuracy	FLASH (kB)	Latency (ms)	RMSE	FLASH (kB)
MONAQ	127.260	**0.746**	257.742	24.729	**9.902**	10.582
w/o Query Rewriting	206.871	0.651	17.186	14.623	11.994	10.155
w/o $\mathcal{A}_{design}$	863.358	0.647	518.762	95.654	13.512	33.243
w/o $\mathcal{A}_{eval}$	540.411	0.641	4775.661	26.335	12.783	109.627
w/o $\mathcal{A}_{eval}$ & $\mathcal{A}_{search}$	601.313	0.643	5907.363	188.123	11.261	665.110
Only $\mathcal{A}_{code}$	579.876	0.612	4158.205	21.530	12.274	99.638

Table 3. Ablation study results of multimodal query generation and multi-agent based search components.

Agents	Query Modality			Classification			Regression		
	Numerical Time Series	Textual Descriptions	Time Series Images	Latency (ms)	Accuracy	FLASH (kB)	Latency (ms)	RMSE	FLASH (kB)
Single (GPT-4o Backbone)	✓			519.159	0.679	3349.453	23.797	13.944	125.445
		✓		1017.267	0.665	4126.024	42.779	13.227	134.901
			✓	593.541	0.690	5971.792	35.859	12.562	193.926
	✓	✓		807.459	0.628	4926.157	40.485	13.556	137.581
	✓	✓	✓	557.665	0.629	3871.910	22.726	12.681	90.398
Multiple (GPT-4o Backbone)	✓			149.320	0.434	**12.066**	54.270	12.284	10.611
		✓		170.461	0.440	15.198	110.751	12.084	12.560
			✓	280.198	0.661	15.638	**13.661**	11.653	**7.885**
	✓	✓		205.623	0.517	16.035	28.049	13.207	13.875
	✓	✓	✓	**127.260**	**0.746**	257.742	24.729	**9.902**	10.582

Hyperparameter Study