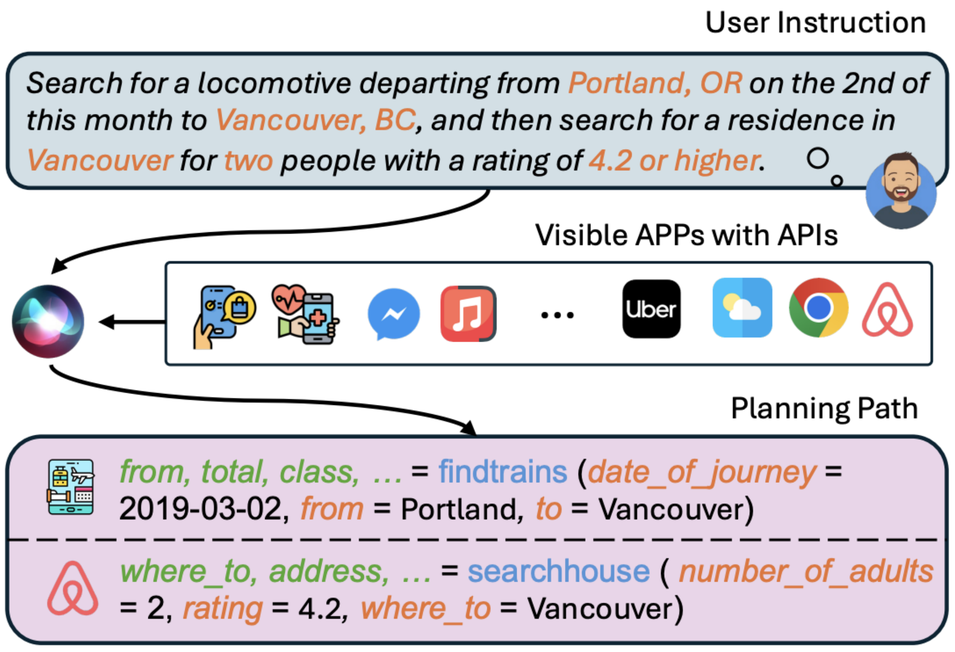
AppBench: Planning of Multiple APIs from Various Apps for Complex User Instruction

Hongru Wang, Rui Wang, Boyang Xue, Heming Xia, Jingtao Cao, Zeming Liu, Jeff Z. Pan, Kam-Fai Wong
The Chinese University of Hong Kong, University of Edinburgh
Hong Kong Polytechnic University

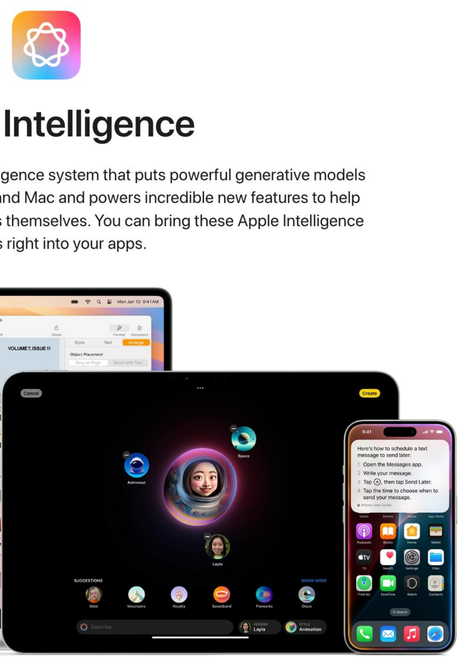
Seek Postdoc Positions

Do check my CV here!

❑ How to evaluate Apple Intelligence to call multiple APIs from various Apps?



App
APIs
Arguments



❑ Two key Challenges and Four Cases in Real-World

Data Type	Example	Structure
SS	Instruction: *Find a house with a rating of 4.6 or higher for a trip to Delhi for two people, inquire about laundry service availability* Output: House: address, phone_number, total_price, has_laundry_service, … = searchhouse(number_of_adults='2', rating='4.60', where_to='Delhi')	Para.=1 Seq.=1
SM	Instruction: *Please book a Hatchback car with insurance to be picked up from Warsaw Chopin Airport on March 7th at 1:30 pm, and returned on March 13th in Warsaw.* Output: Rents: pickup_location, price_per_day, ... = getcarsavailable(car_type='Hatchback', city='Warsaw', end_date='2019-03-13', pickup_time='13:30', start_date='2019-03-07') Rents: car_type, car_name, ... = reservecar(add_insurance='True', car_type=car_type, end_date=end_date, pickup_location=**#pickup_location**, pickup_time=pickup_time, start_date=start_date)	Para.=1 Seq.=2
MS	Instruction: *Search for a locomotive departing from Portland, OR on the 2nd of this month to Vancouver, BC, and then search for a residence in Vancouver for two people with a rating of 4.2 or higher.* Output: Train: from, total, class, … = findtrains (date_of_journey = 2019-03-02, from = Portland, to = Vancouver) House: address, phone_number, total_price, has_laundry_service, … = searchhouse(number_of_adults='2', rating='4.2', where_to= Vancouver')	Para.=2 Seq.=(1,1)
MM	Instruction: *Please make a reservation for 3 people at one Korean restaurant in San Francisco at 1:30 pm on March 12th, and also book a Luxury taxi for 3 to 4 Embarcadero Center.* Output: Restaurant: restaurant_name, has_vegetarian_options, phone_number, rating, address, price_range, category, ... = findrestaurants (category ='Korean', has_seating_outdoors='True', location='San Francisco') Restaurant: date, time, location, ... = reserverestaurant (date='2019-03-12', location=location, number_of_seats='3', restaurant_name = **#restaurant_name**, time='13:30') Rents: destination, ride_type, ride_fare, wait_time, number_of_seats = getride(destination='4 Embarcadero Center', number_of_seats='3', ride_type='Luxury')	Para.=2 Seq.=(2,1)

Two Key Challenges

- **Graph structure**: some APIs can be executed independently while others need to be executed one by one, resulting in graph-like execution order;
- **Permission constraints**: which source is authorized to execute the API call.

Four Cases

- **Single App Single API (SS):** common cases
- **Single App Multiple API (SM):** mostly *sequential*
- **Multiple App Single API (MS):** mostly *parallel*
- **Multiple App Multiple API (MM):** both *sequential* and *parallel*

❑ Evaluation on existing LLMs, and does in-context learning or finetuning help?

Models	SS			SM			MS			MM		
	$F1_{app}$	$F1_{api}$	Succ	$F1_{app}$	$F1_{api}$	Succ	$F1_{app}$	$F1_{api}$	Succ	$F1_{app}$	$F1_{api}$	Succ
Mistral-7B	55.97	16.31	0.51	36.59	15.09	0.50	33.72	6.42	0.00	28.92	7.56	0.00
Vicuna-13B	43.20	3.70	2.00	34.71	4.63	0.50	20.43	3.10	0.00	21.05	2.52	0.00
LLaMA3-8B	63.04	42.67	23.23	37.20	25.33	0.50	30.65	19.52	0.10	26.39	17.80	0.05
LLaMA3-70B	71.20	70.00	50.00	46.48	46.96	10.50	32.61	32.96	2.50	28.97	28.53	0.50
QWen1.5-7B	48.14	19.54	0.00	30.13	16.71	0.00	23.24	10.11	0.00	23.76	11.55	0.00
QWen1.5-14B	72.89	28.41	10.10	41.89	25.51	1.50	42.22	21.98	0.80	32.36	15.07	0.00
QWen1.5-72B	81.23	24.28	12.50	**51.89**	25.27	1.00	**45.94**	13.42	0.62	**38.53**	11.51	0.00
GPT-3.5	63.60	57.95	30.81	41.49	43.65	6.50	33.17	34.53	7.00	27.79	28.09	1.00
GPT-4o	**88.31**	**86.87**	**70.92**	50.83	**50.57**	**20.50**	39.39	**39.14**	**11.00**	32.62	**32.35**	**2.00**

NO!

Explore the original paper for in-depth insights and comprehensive analysis!

Hongru is looking for postdoc position starting from **Aug 2025**.
Research Interests: Tool learning (cognitive tool and physical tool), Large Language Models and Agents, Personalization, Reasoning, …