Beyond Static Evaluation: A Dynamic Approach to Assessing AI Assistants' API Invocation Capabilities

Honglin Mu, Yang Xu, Yunlong Feng, Xiaofeng Han
Yitong Li, Yutai Hou, Wanxiang Che

Abstract

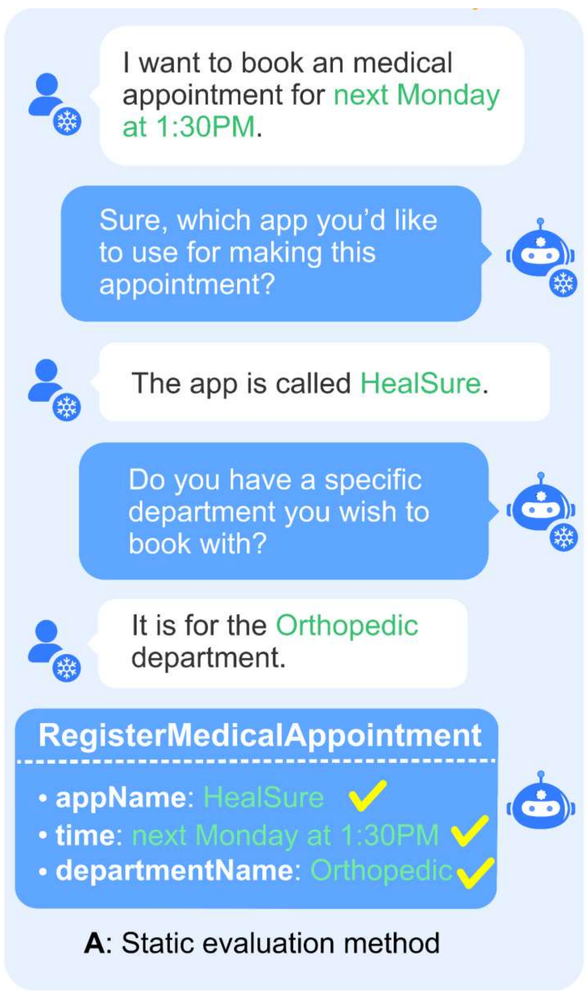
The rapid advancement of Large Language Models (LLMs) has significantly improved AI assistants' capabilities, especially in making API calls. This progress demands more accurate evaluation methods. Currently, many studies assess AI assistants' effectiveness in API calls through static evaluation—using a fixed dialogue history to measure the accuracy of these calls.

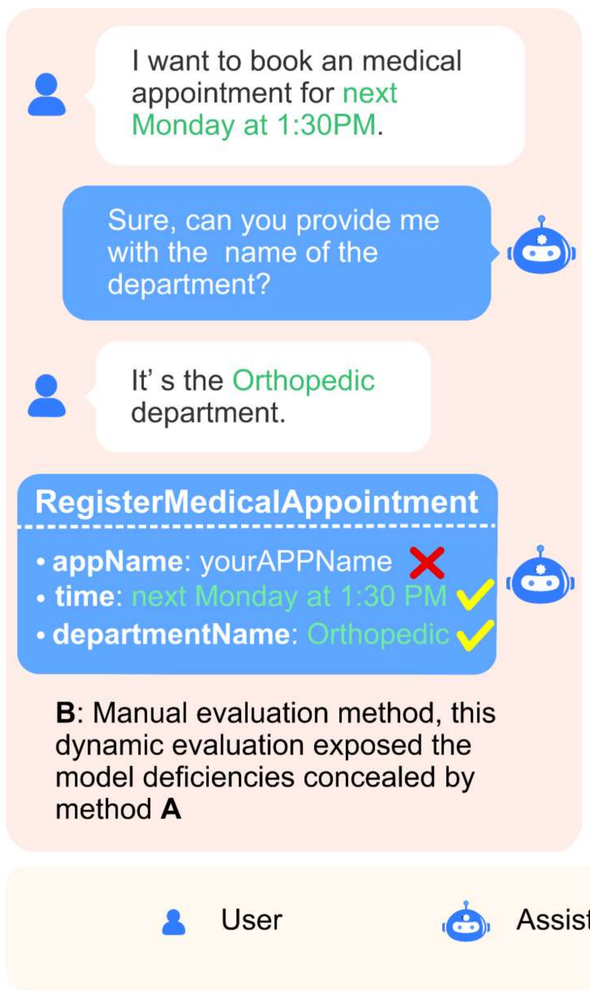
A: During evaluations, the AI assistant does not participate in live dialogue but uses a predetermined dialogue history to make API calls.

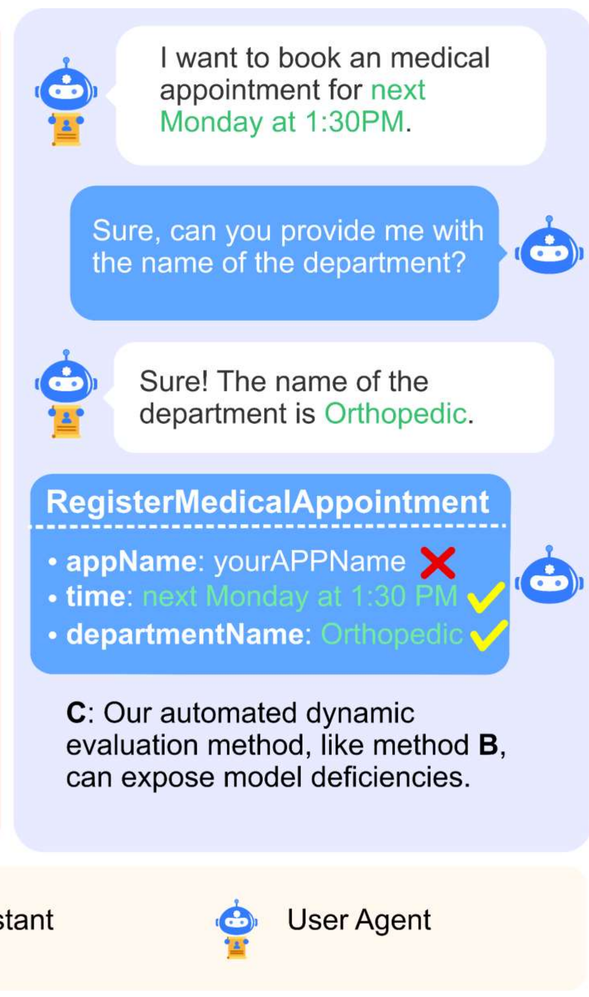
B: However, static evaluation methods may overestimate a model's capabilities by not accounting for issues like missing information and hallucinations in real human-AI interactions.

C: We propose an Automated Dynamic Evaluation (AutoDE) method, which allows for a more accurate assessment of an assistant's API call performance without human involvement.

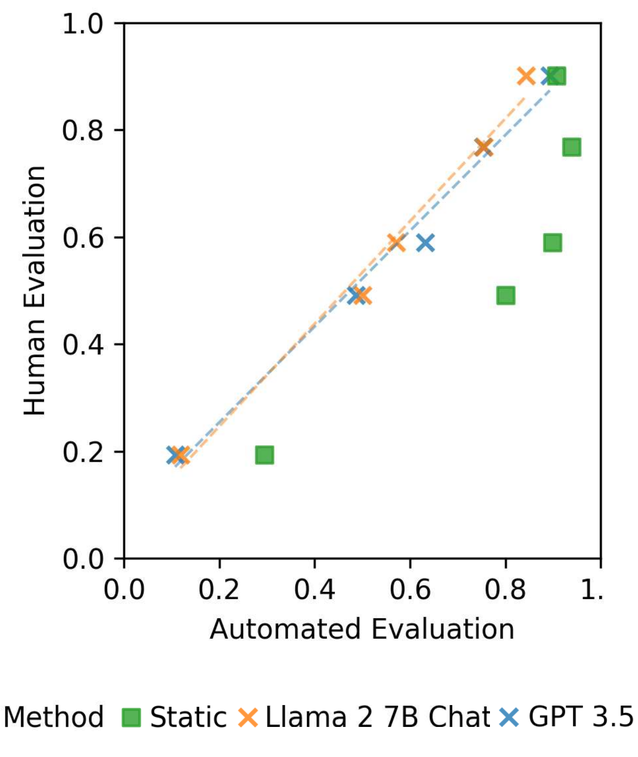
Experimental results highlight that AutoDE uncovers errors overlooked by static evaluations, and further mirrored human evaluation compared to conventional static evaluations.



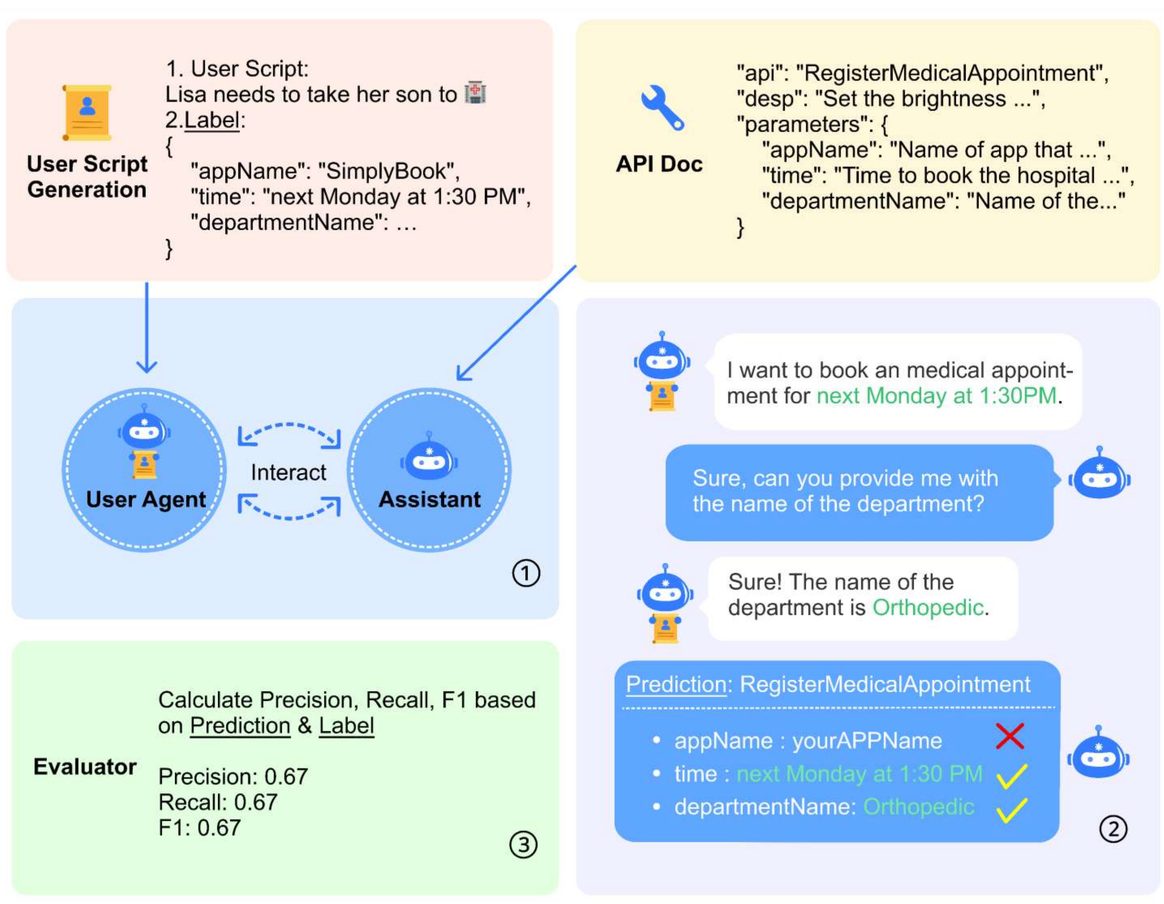




Eval Method	ICC3	R
GPT 3.5	0.9869	0.9923
Llama 2 7B Chat	**0.9923**	**0.9930**
Static	0.8813	0.8813



Automated Dynamic Evaluation



① Our approach, AutoDE, aims to design an automated dynamic evaluation mechanism that approximates human evaluation processes.

② We introduce language models as user agents to simulate human annotators' behavior and guide AI assistants in calling APIs through interactive rounds.

③ The accuracy of API calls generated by the AI assistant is evaluated against expected calls, assessing its tool invocation ability.

Assistant	GPT 3.5			Llama 2 7B Chat			Static			Human		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
GPT 3.5	78.14	73.84	$75.43_{\pm 0.56}$	78.25	73.91	$75.47_{\pm 0.46}$	**94.05**	**93.80**	$\mathbf{93.86_{\pm 0.97}}$	79.62	75.07	76.77
Claude	**91.20**	**88.49**	$\mathbf{89.33_{\pm 1.72}}$	**86.32**	**83.69**	$\mathbf{84.38_{\pm 0.64}}$	93.28	89.53	$90.78_{\pm 0.96}$	**92.60**	**88.74**	**90.05**
Code Llama	64.00	64.41	$63.21_{\pm 3.28}$	56.70	59.30	$57.10_{\pm 2.25}$	91.18	89.74	$89.90_{\pm 0.55}$	59.46	59.99	58.97
Llama Chat	10.61	11.06	$10.71_{\pm 2.32}$	11.48	12.92	$11.86_{\pm 1.41}$	29.30	29.78	$29.40_{\pm 1.61}$	18.80	20.63	19.40
Zephyr	48.08	50.16	$48.69_{\pm 1.68}$	49.76	51.21	$50.05_{\pm 2.39}$	80.69	79.77	$80.01_{\pm 1.92}$	48.70	50.26	49.14

Data Construction

- Constructed APIs for voice assistants:

```
{
    "domain": "Device Manipulation",
    "subdomain": "Setting Item",
    "function": "Luminance",
    "api": "SetLuminance",
    "desp": "Set the brightness ...",
    "parameters": {
        "deviceType": "Supported
            device types ...",
        "targetValue": "Target
            brightness size"
    }
}
```

- The generated user scripts:

```
Character: Lisa, a busy mother
Background: Lisa needs to take her son,
    who recently fell and sprained his
    ankle, to the orthopedic department.
Purpose: Using a tablet, Lisa books an
    appointment at the hospital using a
    medical appointment registration app.
{
    "funcName": "RegMedAppt",
    "time": "Monday",
    "departmentName": "Orthopedic"
}
```

- The user agent prompt:

```
You are an experienced data annotator.
You need to act as a user in a set of
conversations between a user and a voice
assistant Bob ...

Please construct user queries or res-
ponses according to the following
settings:
{{USER_SCRIPT}}
```

LREC-COLING 2024