ShortcutsBench: A Large-Scale Real-World Benchmark for API-based Agents

Haiyang Shen, Yue Li, Desong Meng, Dongqi Cai, Sheng Qi,

Li Zhang, Mengwei Xu, Yun Ma*

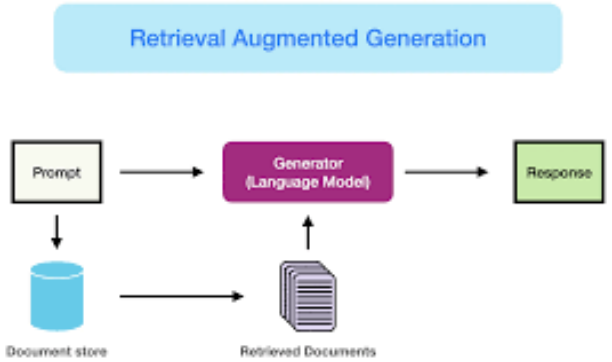
The Thirteenth International Conference on Learning Representations

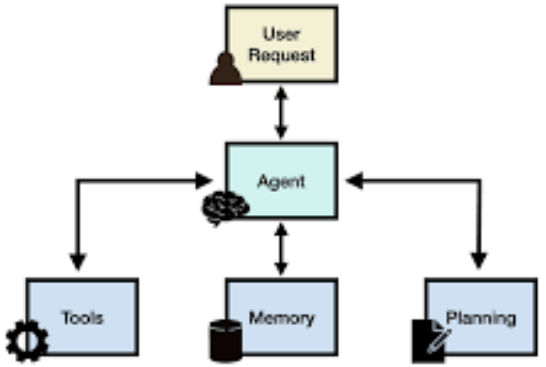
北京大学
PEKING UNIVERSITY

北京郵電大學
BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

Introduction: The Rise of LLM-Based Agents

- The Rise of LLM-Based Agents
  - LLM-based agents are rapidly gaining popularity (academia & industry).
  - Agents leverage APIs to:
    - Access real-time information.
    - Reduce hallucinations with external knowledge.
    - Plan and execute complex, multi-step tasks.
  - Impressive performance on simple tasks.





The Big Question: Beyond Simple Tasks

- Can Current Agents Handle Real-World Complexity?
- **Central Question:** Are API-based agents *truly* capable of handling complex, real-world user demands?

Limitations of Existing Benchmarks:
Part 1 - Lack of Richness and Complexity

- **Limited API Richness:** Few APIs, small number of apps, narrow range of task difficulties.
- **Low Query Complexity:** Average action length is short (1-5.9).
- **Consequence:** Fails to differentiate the capabilities of different LLMs, even less intelligent ones.

Less Intelligent LLMs (even 3B) on existing benchmarks/dataset demonstrated excellent results

Model	LLaMA-3.2-3B	QWen-2.5-3B	LLaMA-3-8B	QWen-2.5-7B	GPT-4o-mini
Acc. MetaTool	89.64	88.29	89.00	92.50	88.31
Acc. ToolLLM	72.92	77.86	78.31	82.69	84.50
Acc. ToolBench	79.47	91.35	93.57	94.26	89.90

Limitations of Existing Benchmarks:
Part 2 - Lack of Realism and Missing Elements

- **Lack of Realism:** Manually crafted APIs, queries not reflecting real user demands.
- **Incomplete Evaluation:**
  - Focus primarily on API selection.
  - Ignore parameter filling (crucial for task completion).
  - Don't assess the ability to request missing information from users/systems.

Introducing SHORTCUTSBENCH: A New Benchmark

- **SHORTCUTSBENCH:** A Realistic, Rich, and Comprehensive Benchmark
- **Key Idea:** Leveraging *real data* from Apple Shortcuts (a Digital Automation Platform - DAP).
- **Advantages:**
  - **Real APIs:** Extracted directly from a widely used platform.
  - **Real Queries:** Reflect actual user needs and a wide range of complexities.
  - **High-Quality Action Sequences:** Human-annotated for accuracy.
  - **Precise Parameter Filling:** Includes primitive types, enums, and output from previous actions.
  - **Evaluation of Missing Information Requests:** Assesses the agent's ability to handle incomplete queries.

SHORTCUTSBENCH: Key Features and Advantages

- Why SHORTCUTSBENCH Stands Out?

Resource	Shortcuts Bench (Ours)	Meta Tool 2024b	Tool LLM 2024	API Bench 2024	Tool Alpaca 2023	API Bank 2023	Tool Bench 2024	Tool QA 2024	Tool Lens 2024
Real API?	✓	✓	✓	✓	✓	✗	✗	✗	✓
Demand-driven Query?	✓	✗	✗	✗	✗	✗	✗	✗	✗
Human-Annotated Act.?	✓	✗	✗	✗	✗	✗	✗	✗	✗
Multi-APIs Query?	✓	✓	✓	✗	✗	✗	✗	✓	✓
Multi-Step Act.?	✓	✓	✓	✗	✓	✓	✓	✓	✓
Prec. Val. for Para. Fill?	✓	✗	✗	✗	✗	✗	✗	✗	✗
Awareness for Ask Info?	✓	✗	✗	✗	✗	✗	✗	✗	✗
# Apps	88	N/A	3451	3	N/A	N/A	8	N/A	N/A
# APIs	1414	390	16464	1645	53	400	232	13	464
# Queries	7627	21112	12657	17002	3938	274	2726	1530	18770
# Avg APIs	9.62	1.02	2.3	1.0	1.0	2.1	5.4	3.5*	2.65
# Avg Actions	21.62	1.02	4.0	1.0	1.0	2.2	5.9	3.9*	2.67

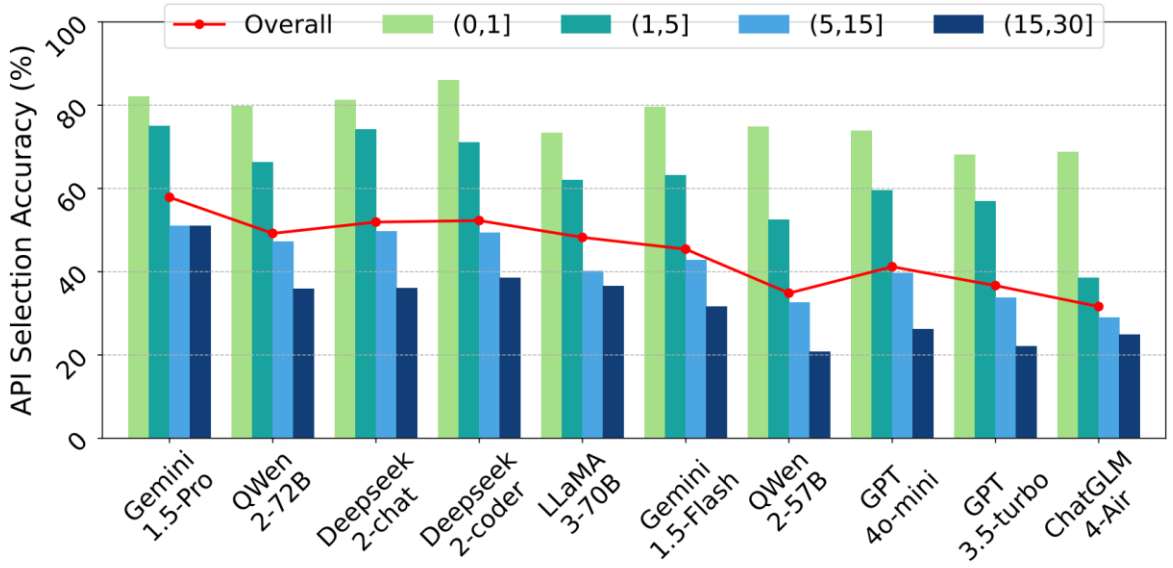
* denotes estimation.

Evaluating Leading LLMs on SHORTCUTSBENCH

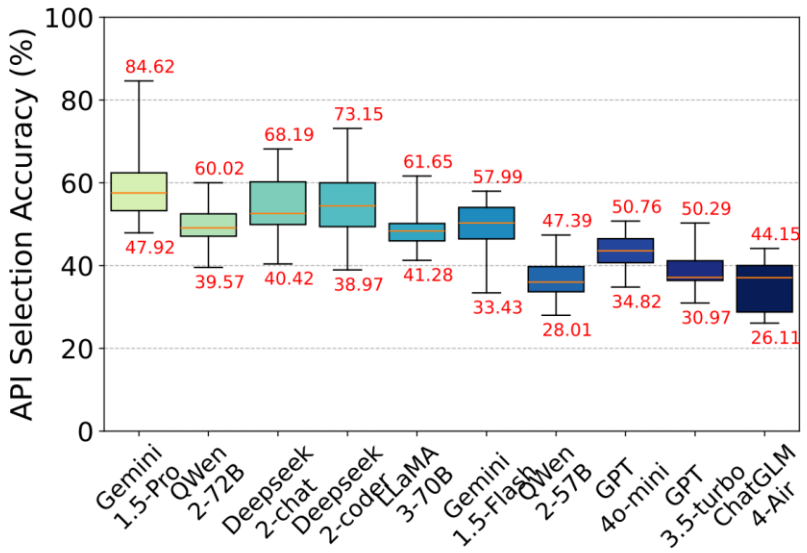
- Rather Comprehensive Evaluation of **21 Leading LLMs**
- Evaluated **both open-source and closed-source** LLMs.
- Covered a **wide range of intelligence levels**.
- Assessed **all key aspects** of agent performance:
  - API Selection
  - Parameter Value Filling
  - Recognition of Missing Information Needs
- **Fully** open-sourced all the datasets, code, experimental logs, and results, and provided detailed documents:
  - https://github.com/EachSheep/ShortcutsBench

SHORTCUTSBENCH Reveals Performance Gaps: API Selection

- **API Selection:** Open-Source Catching Up, but Gaps Remain
- More Results and Analysis in Appendix.



The API selection accuracy on queries with different complexity levels.



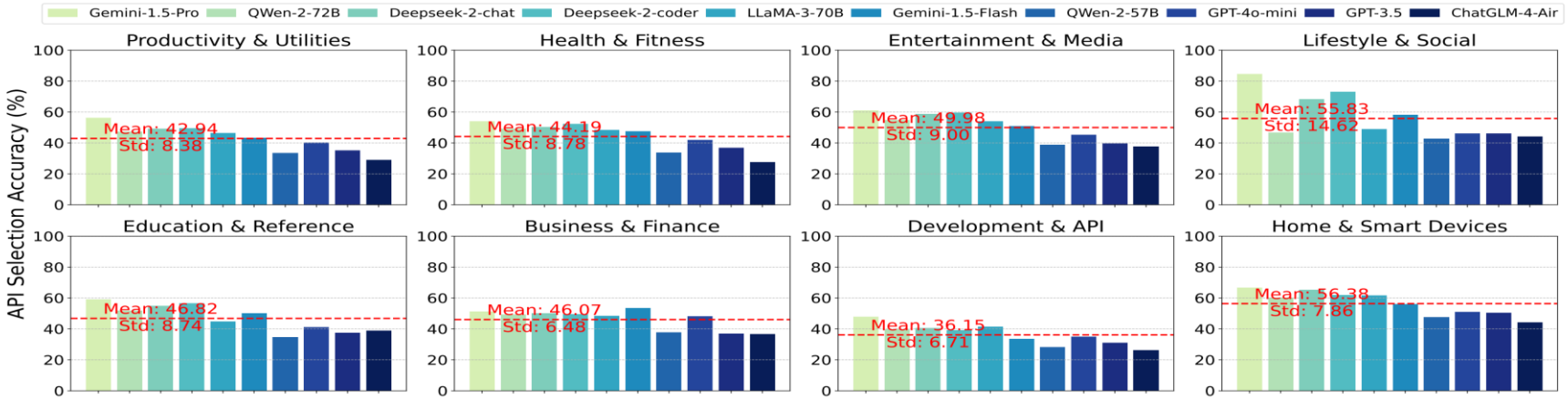
The API selection accuracy difference of each LLM across 8 task types.

SHORTCUTSBENCH Reveals Performance Gaps: API Selection

- **API Selection:** Open-Source Catching Up, but Gaps Remain
- More Results and Analysis in Appendix.
  - **Open vs. Closed Source:** Open-source (>= 70B) matches closed-source on *lower* difficulty, but lags on *higher* difficulty. (This shows the benchmark's ability to distinguish).
  - **Difficulty Matters:** Accuracy drops significantly as task difficulty increases, even for top models like Gemini-1.5-Pro. (Again, highlights differentiation). Quantify the drop (19% and 46%).
  - **LLM-Specific Variations:** Agents built with the *same* LLM show performance differences across task types. (Shows the benchmark's sensitivity). Give the range (15.94% to 36.70%).

Performance Varies by Task Domain

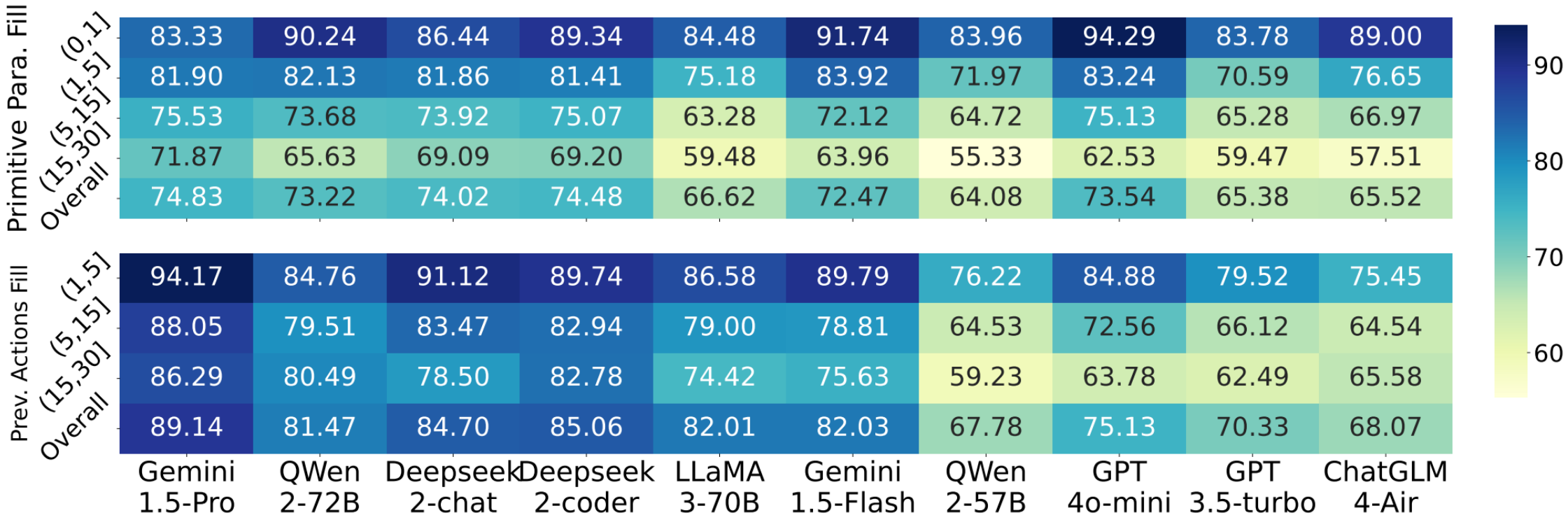
- Task Domain Impacts Performance
  - Agents excel in "Lifestyle & Social" tasks.
  - Performance is significantly lower on "Development & API" tasks (by ~18%).
  - **Implication:** Highlights the need for benchmarks that cover diverse, real-world scenarios.



The API selection accuracy of each task type on 10 API-based agents.

Parameter Filling: A Deeper Dive

- **Parameter Filling:** Accuracy and Challenges



Parameter Filling: A Deeper Dive

- **Parameter Filling:** Accuracy and Challenges
  - **API Selection vs. Parameter Filling:** API selection has a greater impact on overall performance for more intelligent LLMs. (This is a key insight about where the bottleneck lies).
  - **Less Intelligent LLMs Struggle:** Parameter filling accuracy drops significantly with increasing difficulty for less intelligent models.
  - **Extraction is Harder:** Extracting parameters from the user query is more challenging than using previous action outputs. Quantify the accuracy drop (2.55% to 15.39%).

The Missing Piece: Recognizing Information Needs

- **A Critical Weakness:** Failing to Ask for Help

Levels	Gemini 1.5 Pro	QWen 2 72B	Deep seek2 chat	Deep seek2 coder	LLaMA 3 70B	Gemini 1.5 Flash	QWen 2 57B	GPT 4o mini	GPT 3.5 turbo	Chat GLM4 Air
(0, 1]	33.33	37.78	64.29	62.71	47.62	62.79	22.22	37.14	28.89	47.62
(1, 5]	45.95	50.40	55.50	60.08	44.08	53.99	37.24	40.55	37.70	48.06
(5, 15]	51.85	36.42	40.76	49.44	35.71	40.65	28.37	29.71	20.33	48.42
(15, 30]	46.67	25.00	27.59	43.14	22.22	44.64	8.11	38.89	17.14	48.89
Overall	46.59	41.97	47.90	55.18	49.89	40.71	30.74	36.71	30.55	48.28

The Missing Piece: Recognizing Information Needs

- **A Critical Weakness:** Failing to Ask for Help
  - All agents, regardless of size, perform poorly in recognizing when they need more information.
  - Even larger LLMs (like Deepspeed-2-chat 236B) show limited improvement.
  - **Implication**: A major obstacle for real-world deployment, where user queries are often incomplete or ambiguous.

SHORTCUTSBENCH - A Superior Benchmark

- **Realism**: Built on real-world data from Apple Shortcuts.
- **Richness and Complexity**: Covers a wide range of APIs, queries, and difficulty levels.
- **Comprehensive Evaluation**: Assesses API selection, parameter filling, and missing information recognition.
- **Differentiates LLMs**: Clearly shows performance differences between LLMs of varying intelligence.
- **Reveals Key Limitations**: Highlights the challenges in reasoning, parameter extraction, and proactive interaction.ng Progress in LLM-Based Agents
- **Fully** open-sourced all the datasets, code, experimental logs, and results, and provided detailed documents:
  - https://github.com/EachSheep/ShortcutsBench

THANKS!


hyshen@stu.pku.edu.cn