ETOM: A Five-Level Benchmark for Evaluating Tool Orchestration within the MCP Ecosystem

Jia-Kai Dong*, I-Wei Huang*, Chun-Tin Wu, Yi-Tien Tsai

National Taiwan University

EACL 2026 RABAT • MOROCCO
Mars • March 24-29, 2026 • مارس

paper github

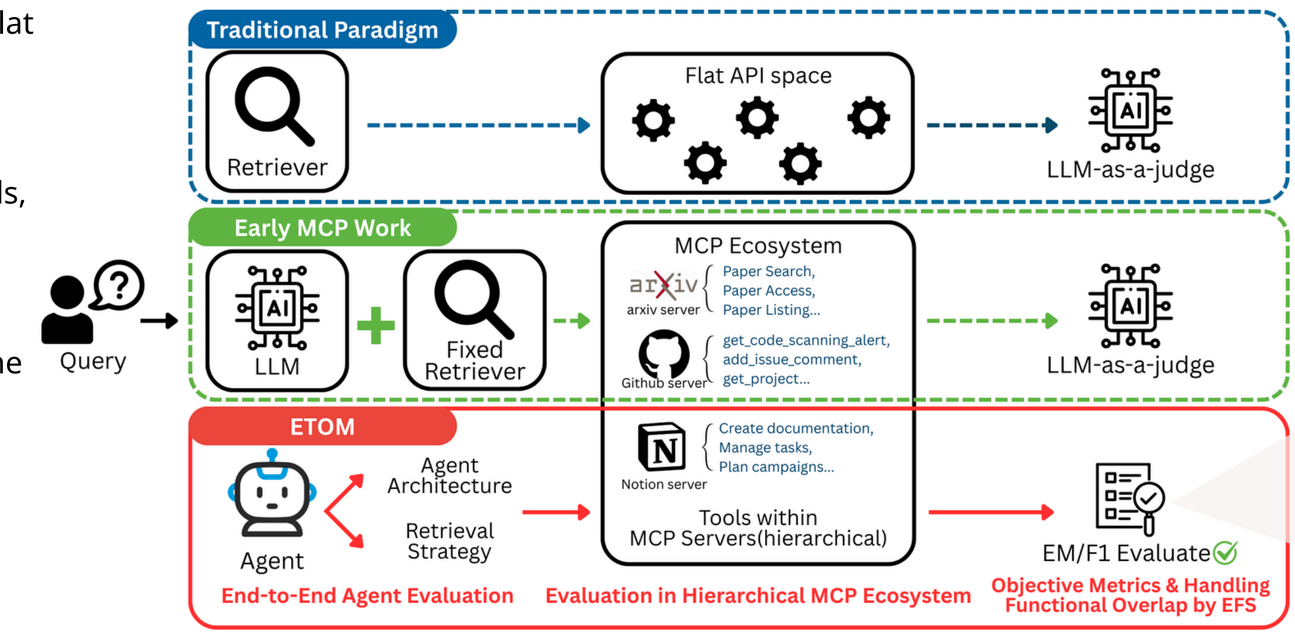
1. Motivation & The Gap

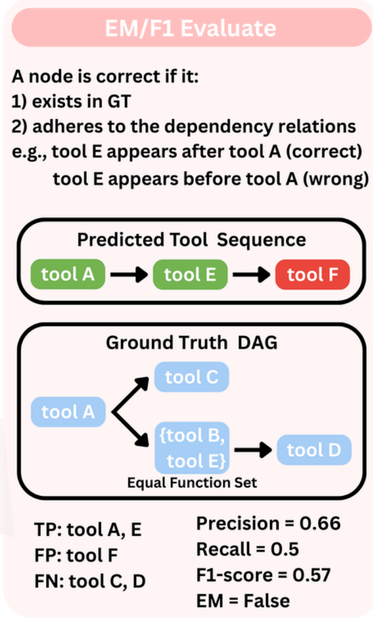
The Paradigm Shift: LLM agents are transitioning from flat API calls to orchestrating tools within distributed Model-Context Protocol (MCP) ecosystems.

The Evaluation Gap: Current benchmarks (e.g., Seal-Tools, NESTful) assume a flat namespace, failing to account for:

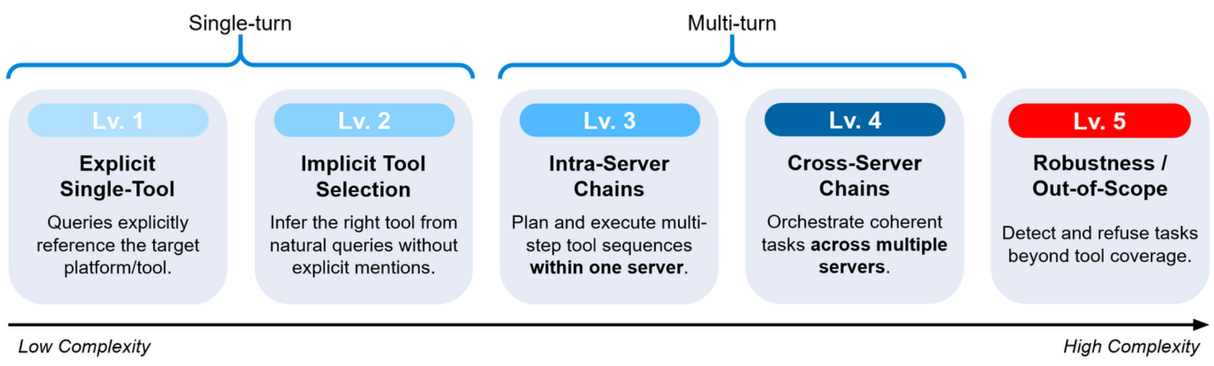
- **Server Boundaries:** The complexity of multi-server orchestration.
- **Functional Overlap:** Multiple tools achieving the same outcome.
- **Robustness:** Identifying out-of-scope user requests.

Our Solution (ETOM): The first end-to-end benchmark evaluating multi-hop orchestration in a realistic MCP environment (491 Servers, 2,375 Tools).





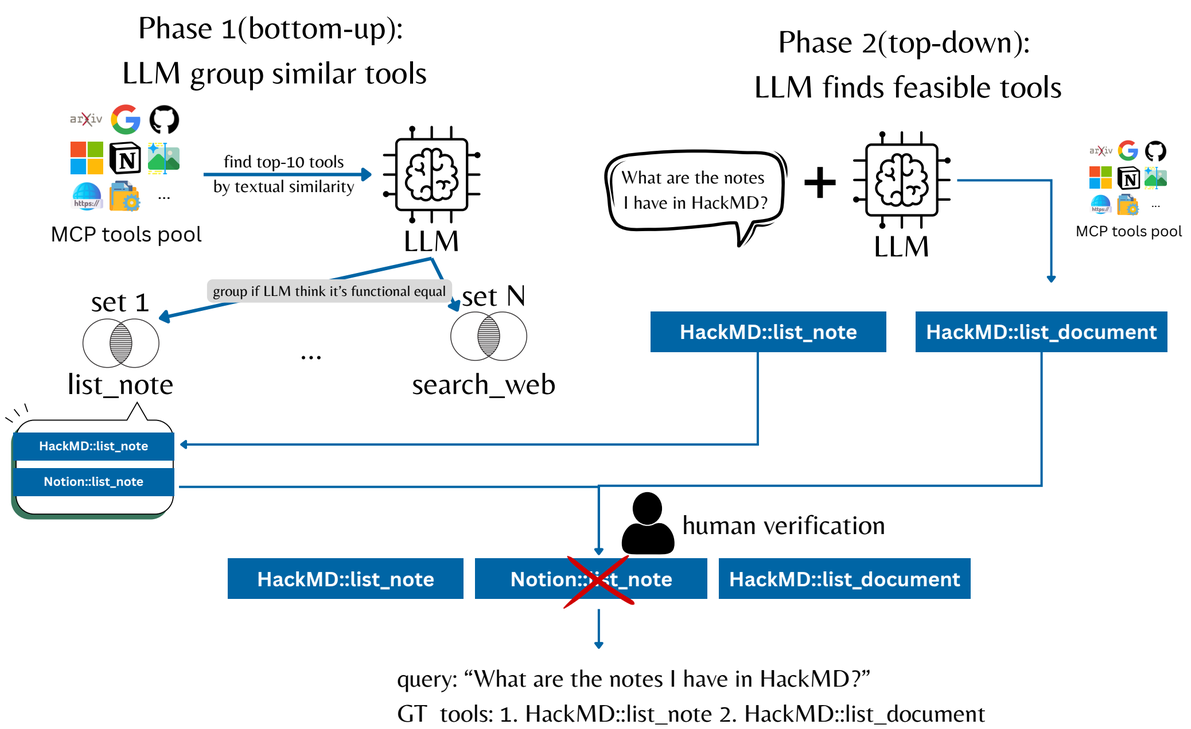
2. ETOM Design & Curriculum



ETOM systematically stress-tests agents from foundational retrieval to complex planning:

- **Levels 1 & 2 (Retrieval):** Tests basic tool discovery. L2 introduces disambiguation among functionally equivalent tools.
- **Level 3 (Intra-Server):** Requires agents to plan sequential chains within a single server.
- **Level 4 (Cross-Server):** The core challenge. Agents must orchestrate compositional workflows across multiple, distinct servers.
- **Level 5 (Robustness):** Tests an agent's ability to reject out-of-scope requests (e.g., physical world interactions).

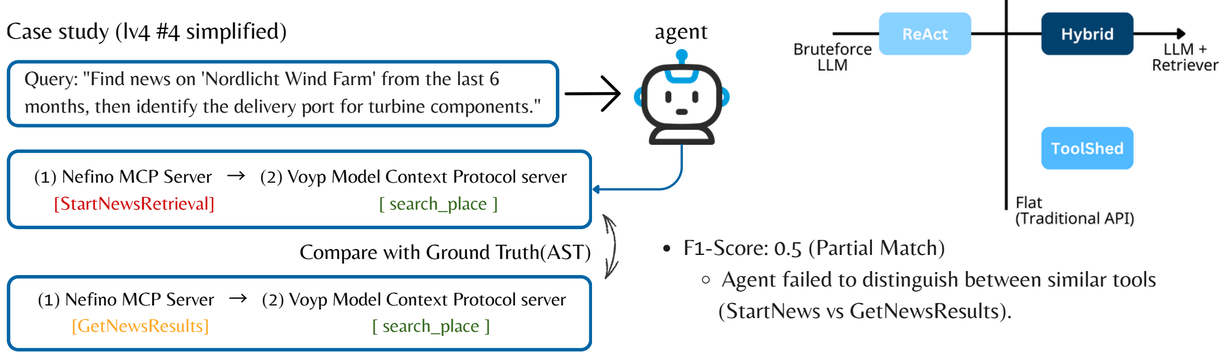
3. Equal Function Sets (EFS)



- **Handling Overlap:** We resolve functional overlap by grouping equivalent tools into Equal Function Sets.
- **Objective Metrics:** EFS enables deterministic, reproducible evaluation (EM/F1 metrics) without relying on costly, biased LLM-as-a-judge pipelines.

4. Experiments & Analysis

Setup: We evaluate 4 agent architectures (ReAct, ToolShed, MCP-Zero, Hybrid) across varying LLM backbones (Llama-3, GPT-4.1).
Task: Given a user query, the agent must retrieve relevant tools from the 2,375-tool corpus and orchestrate a valid execution plan (DAG).



5. Results & Discussion

Method	Backbone	L1 (EM)	L2 (EM)	L3 (F1)	L4 (F1)	L5 (EM)
MCP-Zero	Qwen3-4B	69.0	43.0	46.9	30.4	74.7
ReAct	Qwen3-4B	41.4	19.8	33.6	**6.7**	72.5
ToolShed	Qwen3-4B	**72.7**	51.9	69.4	37.5	70.3
ToolShed	**GPT-4.1**	69.9	49.0	**82.6**	**55.1**	**81.3**

Table 1. performance across different architecture

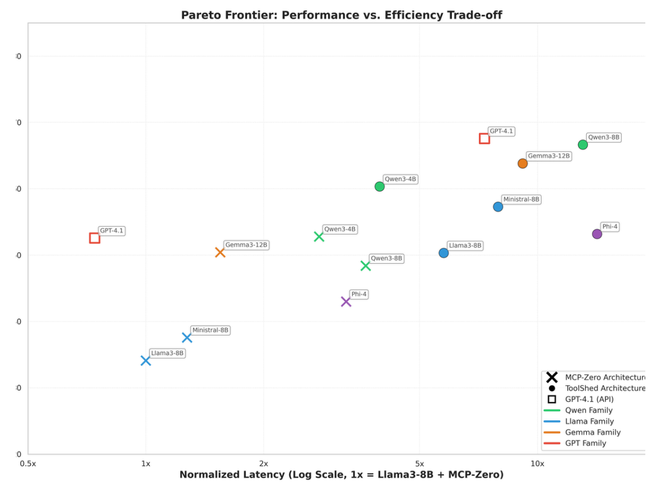


Figure 3. performance-efficiency across different model and arichitecture

Systemic Failure Modes Exposed

- **Contextual Drift:** Agents frequently lose intermediate outputs or explicit constraints when transitioning across server boundaries (L4), decoupling the plan from the user's intent.
- **Premature Decomposition:** Agents tend to over-segment simple queries, introducing hallucinated dependencies and unnecessary failure points.
- **The Safety Gap (L5):** Current frameworks lack architectural safeguards for out-of-scope requests, relying too heavily on an LLM's intrinsic reasoning.

The Core Trade-off: "Hierarchy Isn't Free"

- Hierarchical pruning (MCP-Zero) delivers up to **5.76x efficiency gains** but incurs a significant accuracy cost in complex orchestration.
- Rigid pre-defined hierarchies can stifle the fluid, associative reasoning required for multi-hop tasks (L3/L4).

6. Conclusion

- **ETOM** provides a rigorous, LLM-judge-free diagnostic framework to expose the limitations of tool-using agents in federated ecosystems.
- **Future Direction:** Effective tool orchestration requires model-architecture co-design, shifting toward "Hierarchy-Aware" reasoning and adaptive retrieval strategies to balance efficiency and accuracy.

Jia-Kai Dong (b11901067@ntu.edu.tw) I-Wei Huang (r14922088@ntu.edu.tw)