System 1 agents: local tool-calling for clinical data without letting an LLM write the call

22 November 2026 · 7 min read · blogs.iictp.org/system-1-agents-local-biomedical-tool-calling

Short answer: “System 1 agents” is a method, in an arXiv preprint dated 16 November 2026, for building local agents that call tools without generating free text. A deterministic compiler turns each tool’s schema and the user request into multiple-choice questions, and a one-pass decision model answers them. On 1,952 requests to 157 biomedical tools the authors report 97.0% correct tool choice or correct refusal and no invented content in any call. This is an arXiv preprint. It has not been peer reviewed, and its numbers are the authors’ own claims.

TL;DR The short version

- 'System 1 agents', an arXiv preprint dated 16 November 2026, call tools without letting an LLM write the call as text.
- A deterministic compiler turns each tool schema and request into multiple-choice questions, and a one-pass decision model answers them.
- The best agent chose the right tool or correctly declined for 97.0% of 1,952 biomedical requests, but it is unreviewed and covers only requests that state their arguments.

Read online, listen to it, or change the text size: https://blogs.iictp.org/system-1-agents-local-biomedical-tool-calling

What problem does it solve?

Agents that analyse sensitive clinical data usually rely on large language models. Cloud models expose patient data. Local models hallucinate, and an invented value in a tool call propagates through every later step. The authors, Eleftherios Tzanis and Michail E. Klontzas, aim for agents that are local, reliable and fast.

How it works

- **Compile.** A deterministic compiler converts each tool’s schema and the request into multiple-choice questions whose options come only from the request and the schema.
- **Decide.** A one-pass decision model answers those questions. It never writes the tool call as text.
- **Why that helps.** If every option is drawn from the request or the schema, the model cannot insert a value that was in neither, which removes the invented-argument failure by construction.

What is new compared with LLM agents

A typical agent asks an LLM to emit a JSON call. Here the language model's freedom is replaced by a constrained choice, so the agent behaves more like a fast classifier. The name echoes fast, automatic "System 1" thinking.

Results reported by the authors

- 1,952 requests across 157 biomedical tools.
- The best System 1 agent chose the right tool, or correctly declined, for 97.0% of requests, ahead of all 18 local LLM-agent settings, where the best reached 95.2%.
- It filled 97.0% of 6,056 stated argument values.
- No System 1 call contained invented content, against 5.5% to 19.7% of LLM calls.
- The framework is open source, per the abstract.

Limitations and open questions

- Preprint, not peer reviewed.
- It covers requests where the arguments are stated in the request; open-ended reasoning and multi-step planning were not evaluated in the abstract.
- The 3% of requests that were wrong still matter in clinical settings; the paper's error analysis was not read.
- Benchmarks of 157 tools may not match your tool set, so test on your own schemas.

How to try the idea

1. List the tools your agent may call and write down each argument's allowed sources: user text, schema enumerations, or defaults.
2. Check which of your tool failures are invented values; if they dominate, a constrained decision step may help.
3. Read the open-source repository's licence and README before depending on it.
4. Keep a refusal path, and log every declined request for review.

What it means for healthcare practitioners

For regulated data, running locally and constraining outputs are both risk controls. The approach does not replace clinical validation, and it does not remove the need for privacy review. See private AI on-premise for regulated industries.

Frequently asked questions

- **Does it use an LLM?** The method avoids generating text for the call; it uses a one-pass decision model over compiled questions.
- **Is it more accurate than LLM agents?** On the authors' benchmark, yes, with 97.0% against a best of 95.2%.
- **Is it validated for clinical use?** No such claim is made in the abstract.

Related reading

- Private AI and on-premise LLMs for regulated industries
- AI agents explained
- What is LLM hallucination?

Sources and currency

Sources were read on 22 November 2026. Products, rules and standards change, so check the linked originals before you rely on any detail.

> Related credential: AI Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. Tzanis E., Klontzas M. E. System 1 agents: a method for local, reliable and fast agentic AI. arXiv:2611.21053. 16 November 2026. arxiv.org/abs/2611.21053.

Published by IICTP, an independent technical credential body, on 22 November 2026. The current version of this post is at https://blogs.iictp.org/system-1-agents-local-biomedical-tool-calling. More guides: blogs.iictp.org.