Tool Routing for Resource-Constrained LLM Agents

Alex Haščík

Supervisor: Mgr. Marek Šuppa

May 15, 2026

What is Tool Routing?

Agentic AI systems have access to a set of tools / functions.

Tool routing = the model's ability to decide:

- *which* tool to call,
- *what arguments* to pass it,
- based on a natural language query.

Example:

User: *"What's the weather in Vienna?"*

✓ Correct: `get_weather(location="Vienna")`

× Wrong: `search_web(query="Vienna weather")`

Challenge: natural language → selecting the **correct** executable tool and submitting it with exactly the right name, parameters, and values.

Goals of the Thesis

1. Reviewing the current state-of-the-art in tool-routing strategies for agentic LLM systems, with a focus on approaches suitable for small, resource-constrained models.
2. Investigating lightweight methods to improve routing accuracy without relying on larger model backends.
3. Exploring evaluation frameworks and benchmarks for assessing the tool-use capabilities of small language models.
4. Conducting an empirical analysis of the trade-offs between model size, routing performance, and task success rates.

What We Did This Semester

- Studied relevant literature.
- Set up local inference environment (Ollama, Llama.cpp) and tested multiple open-source small language models on local hardware.
- Integrated the **BFCL** (Berkeley Function Calling Leaderboard) evaluation framework
- Registered a custom local model (Qwen3-1.7B) endpoint in BFCL, redirecting its OpenAI-compatible client to the local Ollama server.
- Ran first empirical measurement on the `simple_python` category: **49% accuracy across 400 test cases**.

Next Steps

- Explore other benchmarks, such as SkillsBench, ToolBench, ...
- Evaluate multiple models, with multiple parameter sizes.
- Analyse failure patterns such as wrong tool name vs. wrong parameter vs. wrong value type.
- Test lightweight interventions for improving tool calling accuracy:
  - few-shot prompting,
  - chain-of-thought reasoning,
  - tool pre-filtering using embeddings,
  - ...
- Study relevant literature.

References

- S. Yan et al., *Berkeley Function-Calling Leaderboard*. UC Berkeley, 2024.
- X. Li et al., *SkillsBench: Benchmarking How Well Agent Skills Work Across Diverse Tasks*. arXiv:2602.12670, 2026.

Thank you for your attention