

Adaptive Tool Selection for LLM Agent in Task Solving

Master Thesis Final Report

Presenter: Xincheng Chen

Supervisor: Yuchen Xia

Examiner: Prof. Dr. Ing. Michael Weyrich



Contents

- **Introduction**
- **Technical Basics**
- **RQ1: Which Tools are Frequently Necessary when Solving Scientific Tasks?**
- **RQ2: What is Effectiveness of MCP-Enhanced LLMs for Scientific Problem Solving?**
- **Conclusion and Outlook**

Introduction

Introduction

Limitations of Large Language Models (LLMs) in Problem Solving

Fragile Numerical Reasoning

LLMs can perform basic arithmetic but their performance drops on problems that require deep mathematical reasoning.

Rahman et al., *A Fragile Numerical Reasoning*

$$A = \begin{bmatrix} 1 & 2 & -1 & 3 \\ 2 & 4 & -2 & 6 \\ 0 & 1 & 1 & 1 \end{bmatrix}$$

$r1 \times 2 = r2$

Correct Answer:
rank(A) = 2

Consider the following matrix arising from

$$A = \begin{bmatrix} 1 & 2 & -1 & 3 \\ 2 & 4 & -2 & 6 \\ 0 & 1 & 1 & 1 \end{bmatrix}$$

What is the rank of matrix A?

Mathematical Reasoning

Even when the final answer is correct, the reasoning process is often flawed or

Reasoning Failures, arXiv:2502.11574 (2025)

Systematic Numerical Errors

LLMs frequently make systematic arithmetic mistakes that lead to unreliable results in mathematical computation.

Zhang & Graf et al., *Mathematical Computation and Reasoning Errors*, AIME-Con 2025

WORK:

- Checked for linear independence of rows.
- Verified row operations to reduce to row echelon form.

FINAL ANSWER: 3

→ Wrong Answer!

Peters & Chin-Yee, *Generalization Bias in LLM Scientific Summaries*, arXiv:2504.00025 (2025)

Errors in Scientific Summaries

LLMs often generalize and distort scientific information when summarizing research, revealing a lack of reliable scientific understanding.

- LLM lack reliability in solving complex problems

Introduction

Limitations of Large Language Models (LLMs) in Problem Solving

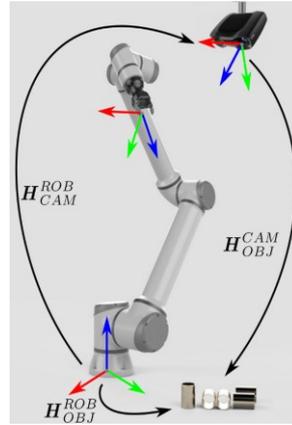
Example: 3-DOF Spatial Robot Forward Kinematics

A 3-DOF robotic manipulator has the following homogeneous transformation matrices:

$$T0^1 = \begin{bmatrix} 0.8660254 & -0.5 & 0 & 0.0866025 \\ 0.5 & 0.8660254 & 0 & 0.0500000 \\ 0 & 0 & 1 & 0.4000000 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$T1^2 = \begin{bmatrix} 0.8660254 & 0.5 & 0 & 0.2598076 \\ -0.5 & 0.8660254 & 0 & -0.1500000 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$T2^3 = \begin{bmatrix} 0.5 & -0.8660254 & 0 & 0.1000000 \\ 0.8660254 & 0.5 & 0 & 0.1732051 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$



Task: Compute $T0^3 = T0^1 * T1^2 * T2^3$ by matrix multiplication and report the end-effector position (x, y, z) rounded to 3 decimals.

Correct answer: (0.487, 0.223, 0.400)

LLM answer:

The screenshot shows a chat interface with the following content:

Chat meta-llama/Meta-Llama-3-8B-Instruct-Lite API view

System Prompt: You are a careful scientific assistant. Output format rul... + >

[[[0.5, -0.8660254, 0, 0.1000000], [0.8660254, 0.5, 0, 0.1732051], [0, 0, 1, 0], [0, 0, 0, 1]]]

Task: Compute $T0^3 = T0^1 * T1^2 * T2^3$ by matrix multiplication and report the end-effector position (x, y, z) rounded to 3 decimals.

WORK:

- Performed matrix multiplication: $T0^3 = T0^1 * T1^2 * T2^3$
- Calculated each product of matrices: $T0^1 * T1^2$, $T0^1 * T1^2 * T2^3$
- Verified the resulting matrix dimensions (4x4) and values

FINAL_ANSWER: $T0^3 = \begin{bmatrix} 0.5 & -0.8660254 & 0 & 0.1732051 \\ 0.8660254 & 0.5 & 0 & 0.1000000 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$;

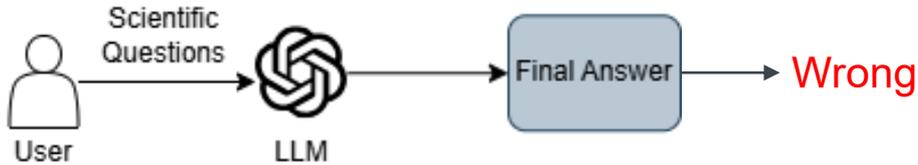
End-effector position (x, y, z) = (0.173, 0.1, 0.4)

(0.173, 0.1, 0.4) → Wrong Answer!

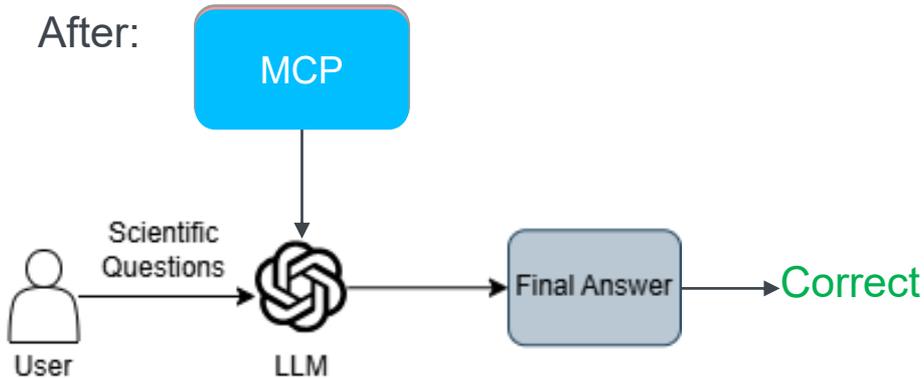
Introduction

How to Support LLMs in Task Solving?

Before:



After:



- **Research Question 1:**

Which Tools are Frequently Necessary when Solving Scientific Tasks?

- **Research Question 2:**

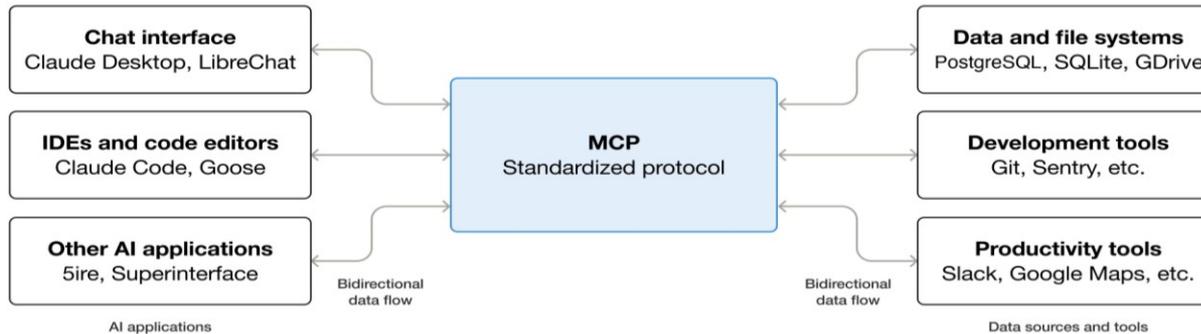
What is Effectiveness of MCP-Enhanced LLMs for Scientific Problem Solving?

Technical Basics

Technical Basics

What is Model Context Protocol (MCP)?

MCP acts as a standardized bridge between LLM-based interfaces and external tools or data sources.

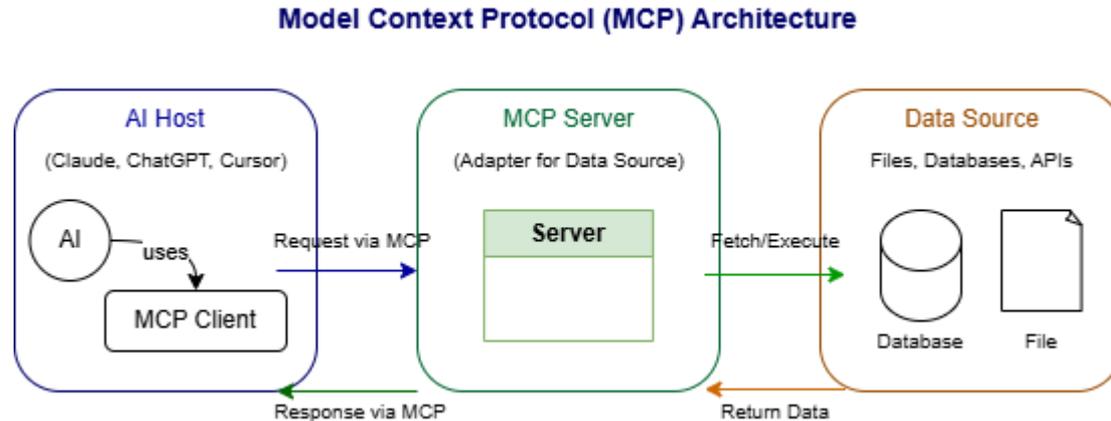


MCP as a bridge enables:

- LLMs **no need** to know how tools are implemented
- Different tools can be accessed through the **same protocol**
- Tools are exposed via **explicit schemas**

Technical Basics

How LLMs Integrate with MCP?



[5] Source: Model Context Protocol (MCP). Official website: <https://modelcontextprotocol.io/>

- *The MCP Client translates AI requests into standardized protocol format, communicates with MCP Servers, which then interact with external data sources*

Research Question 1:

Which Tools are Frequently Necessary when Solving Scientific Tasks?

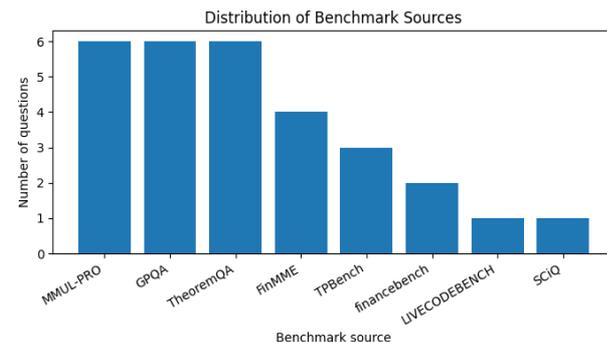
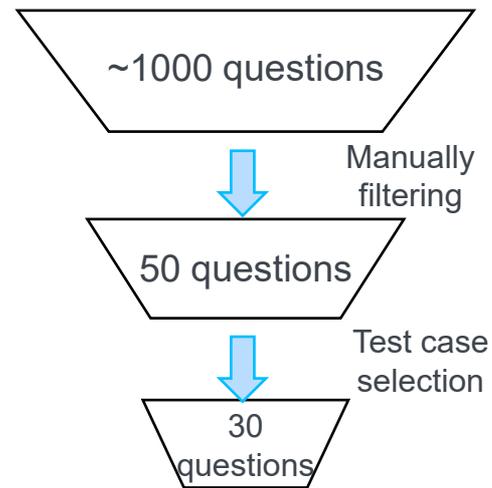
- Benchmark Selection and Question Sampling
- Typical Tasks
- Frequency Analysis of Required Tool Types

RQ1: Which Tools are Frequently Necessary when Solving Scientific Tasks?

Benchmark Selection and Question Sampling

Benchmark	Source	Total tasks
MMLU-Pro		~1,000
MMLU	Question: How many milliliters of 0.250 M KOH does it take to neutralize completely 50.0 mL of 0.150 M H ₃ PO ₄ ? Options: A. 75.0 mL, B. 90.0 mL, C. 60.0 mL, D. 120 mL, E. 30.0 mL, F. 180 mL, G. 270 mL, H. 100 mL, I. 27 mL, J. 150 mL	
SciQ	Answer: B	
TheoremQA	https://github.com/TIGER	~1,000
GPQA	Example 3 Q: Changes from a less-ordered state to a more-ordered state (such as a liquid to a solid) are always what?	~150
TPBench		~100
FinanceBench		~1,000
LiveCodeBench	1) exothermic 2) unbalanced 3) reactive 4) endothermic	~100
FinMME	https://huggingface.co/datasets/luojunyu/FinMME	~1,000
TheoremQA	TheoremQA: Question: Let $W(t)$ be the standard Brownian motion. Find the probability of $P(W(1) + W(2) > 2)$. Wiener's Process The Wiener process W_t is characterised by the following properties: W has independent increment. For every $t > 0$, the future increment $W_{t+u} - W_t$ are independent from the past W_t . W has Gaussian increments, $W_{t+u} - W_t$ has Gaussian distribution $\mathcal{N}(0, u)$. Answer: 0.186, Type: Float	

Randomly sampling



RQ1: Which Tools are Frequently Necessary when Solving Scientific Tasks?

Typical Tasks

Task Example 1

Task Question: Given plant $G(s)$, solve PID gains from target characteristic polynomial. $p_{target}(s) = s^3 + 8s^2 + 23s + 30$. \rightarrow

$$\begin{cases} \frac{3}{2}K_p - K_i + \frac{1}{2}K_d = 23/3 \\ -\frac{5}{3}K_p + \frac{4}{3}K_i - K_d = -10/3 \\ \alpha = 1 \end{cases}$$

Tool Necessary: Math Calculator

MCP Tool Call Example: math_calculator (“linear_system”, “solve”, “[[3/2, -1, 1/2], [-5/3, 4/3, -1], [0, 0, 1]]”, “[23/3, -10/3, 1]”)

Task Example 2

Task Question: At 60°F dew point, find vapor partial pressure and mass fraction.

Tool Necessary: Physics steam table

MCP Tool Call Example: physics_query(“steam_table_iapws97”, “60 °F”)

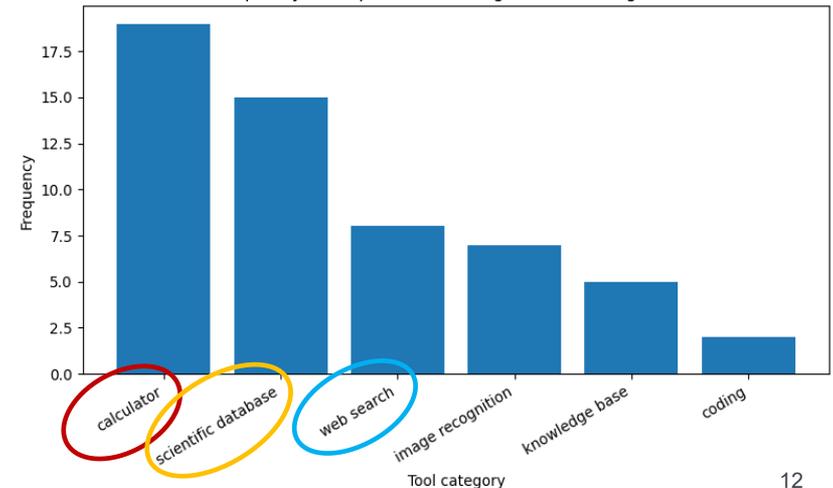
Task Example 3

Task Question: How is the SI meter defined in the post-2019 SI system without a physical artifact?

Tool Necessary: Wikipedia

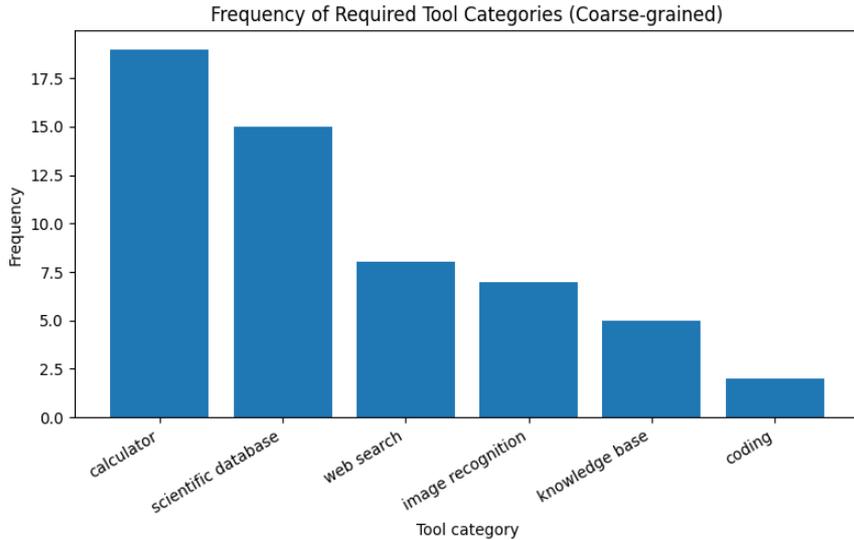
MCP Tool Call Example: wikipedia_query (“wikipedia”, “SI meter definition”)

Frequency of Required Tool Categories (Coarse-grained)



RQ1: Which Tools are Frequently Necessary when Solving Scientific Tasks?

Frequency Analysis of Required Tool Types



MCP
Implementation

- Conclusion: **Calculator**, **scientific databases**, and **web search** are the most frequently required tools for scientific problem solving.
- Limitation: Questions filtering and required tools choosing are based on my subjective judgment

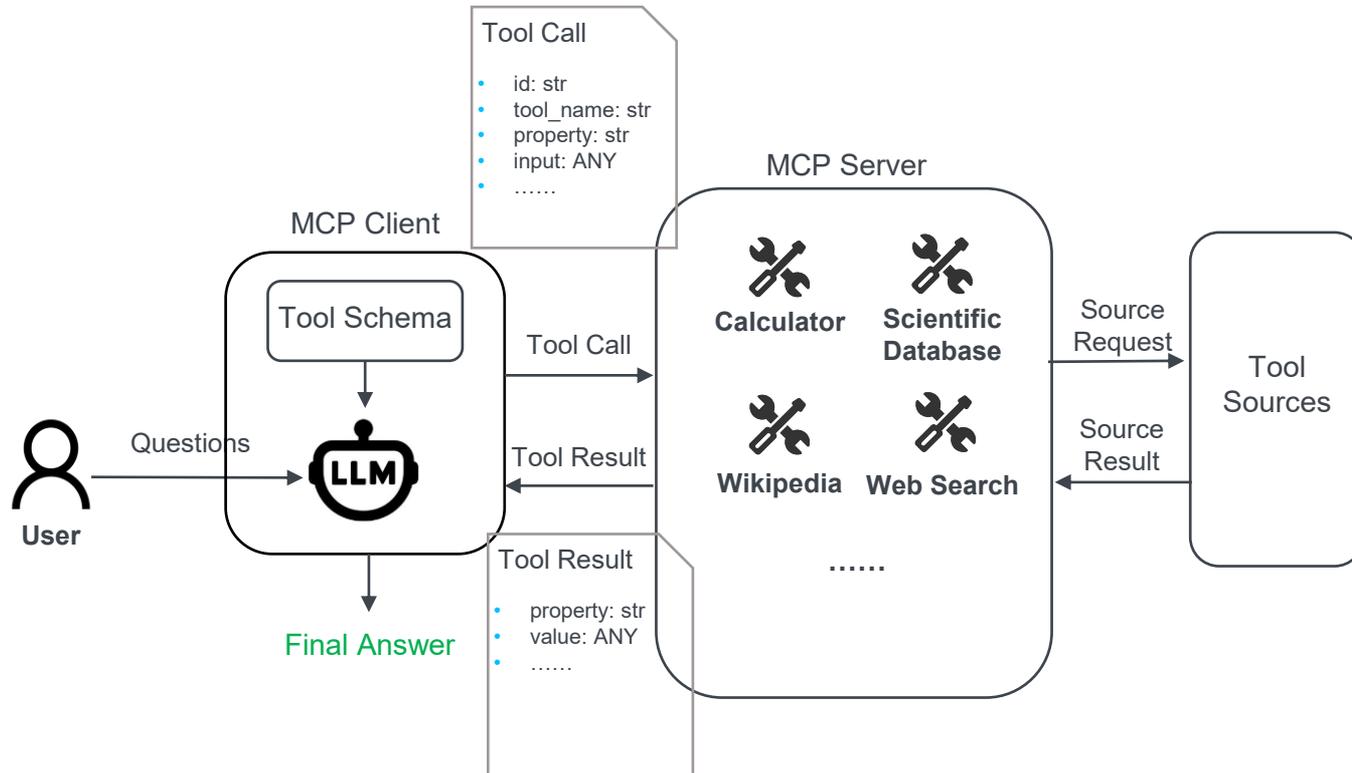
Research Question 2:

What is Effectiveness of MCP-Enhanced LLMs for Scientific Problem Solving?

- MCP-enhanced LLM System Implementation
- Dataset Construction
- Experimental Setup
- Experiment 1: MCP Effectiveness on Small LLM
- Experiment 2: MCP Effectiveness over Multi-models

RQ2: Effectiveness of MCP-Enhanced LLMs

MCP-enhanced LLM System Implementation



RQ2: Evaluation of MCP-Enhanced LLMs

MCP-enhanced LLM System Implementation

Category	Tool Name	Description
External Service Tools	wikipedia	General-purpose knowledge retrieval with Wikimedia
	web_search	Up-to-date or time-sensitive information with Serper
	biology_ncbi_pubmed_esearch	Literature retrieval via NCBI PubMed
	chemistry_pubchem	Retrieval of chemical compound properties from PubChem
	astronomy_solar_system	Planetary and solar system parameters
	engineering_materials_project_elasticity	elastic properties database
	Local Database Tools	physics_steam_table_iapws97
	physics_codata_2022	CODATA 2022 physical constants (local copy)
	chemistry_equilibrium_constants_local	Local database of chemical equilibrium constants
Local Compute Tools	linear_system	Solve multi-unknown linear equations for control and parameter identification tasks.
	matrix_multiply	Compute matrix multiply
	numeric_eval	Perform deterministic numerical evaluation with known variables.
	root_finding & optimization	Unified root solving and optimization for implicit equations and stability analysis.

RQ2

System Demo

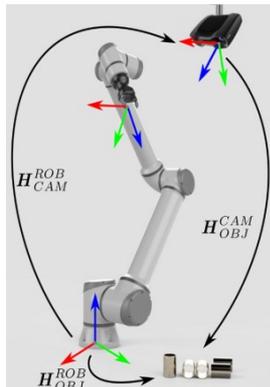
Compute end-effector position from chained homogeneous transforms:

$$T_0^3 = T_0^1 \cdot T_1^2 \cdot T_2^3$$

$$T_0^1 = \begin{bmatrix} 0.8660254 & -0.5 & 0 & 0.0866025 \\ 0.5 & 0.8660254 & 0 & 0.0500000 \\ 0 & 0 & 1 & 0.4000000 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$T_1^2 = \begin{bmatrix} 0.8660254 & 0.5 & 0 & 0.2598076 \\ -0.5 & 0.8660254 & 0 & -0.1500000 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$T_2^3 = \begin{bmatrix} 0.5 & -0.8660254 & 0 & 0.1000000 \\ 0.8660254 & 0.5 & 0 & 0.1732051 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$



MCP Demo

Question

Spatial Robot Forward Kinematics (Numeric Matrices)

A 3-DOF robotic manipulator has the following homogeneous transformation matrices:

$$T_0^1 = \begin{bmatrix} 0.8660254 & -0.5 & 0 & 0.0866025 \\ 0.5 & 0.8660254 & 0 & 0.0500000 \\ 0 & 0 & 1 & 0.4000000 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Ground Truth (GT)

(0.487,0.223,0.400)

Run (MCP)

RQ2: Evaluation of MCP-Enhanced LLMs

Test Dataset Construction consists of 45 Samples



Typical Tasks

Q1 (Robotics / Control)

“An industrial robotic manipulator ... determine constraint matrix rank and closed-loop stability.”

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 6 & 8 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{bmatrix}$$

Q2 (PID / Control)

“A unity-feedback PID control system ... solve gains K_p , K_i , K_d via coefficient matching.”

$$\begin{cases} \frac{1}{2}K_p - K_i + \frac{1}{2}K_d + 2\alpha = \frac{23}{3}, \\ -\frac{5}{3}K_p + \frac{4}{3}K_i - K_d + \frac{2}{3}\alpha = -\frac{10}{3}, \\ \frac{1}{2}K_p + \frac{1}{2}K_i + K_d - \frac{3}{2}\alpha = \frac{11}{12}, \\ \frac{4}{3}K_p - \frac{10}{3}K_i + \frac{8}{3}K_d + \frac{2}{3}\alpha = \frac{34}{3}. \end{cases}$$

Q3 (Physics / Thermodynamics)

“A dew-point beaker ... condensation at 60°F, total pressure 14.7 psia; determine water-vapor partial pressure and mass fraction.”

Q4 (Astronomy)

“Two stars ... given relative Si, Mg, Fe abundances and solar references; compute the ratio of silicon atoms in their photospheres.”

.....

Q45 (Engineering Materials)

“Diamond gasket simulation ... use Materials Project mp-66 VRH K and G to compute compressional modulus $M=K+4/3 \cdot G$; choose best value.”

RQ2: Evaluation of MCP-Enhanced LLMs

Experiment 1: MCP Effectiveness on Small LLM

LLM	Correctness without tool using	Correctness with MCP tool using
Qwen2.5-7B-Instruct-Turbo (small)	31%	47% (+16%)

Evaluation Metrics

- Correctness: Final answer accuracy

$$\text{Correctness} = \frac{\text{Number of correct answered questions}}{\text{Number of total questions}}$$

- MCP significantly improve the performance of small LLMs on scientific task solving

RQ2: Evaluation of MCP-Enhanced LLMs

Experiment 1: MCP Effectiveness on Small LLM (in Depth)

LLM	Correctness without tool using	Correctness with MCP tool using	Tool Usage Frequency	Tool Usage Precision
Qwen2.5-7B-Instruct-Turbo (small)	31%	47% (+16%)	66%	82%

Evaluation Metrics

- Tool Usage Frequency: **when a tool should be used, did the model actually use it?**

$$ToolUsageFrequency_i = \frac{\text{Number of used recommended tools}}{\text{Number of recommended tools}} = \frac{|R_i \cap U_i|}{|R_i|}, R_i > 0$$

R_i = recommended tools for question i
 U_i = tools actually used by MCP for question i

- Tool Usage Precision: **when the model uses tools, are the tools the right ones?**

$$ToolUsagePrecision_i = \frac{\text{Number of actually used tools that are recommended}}{\text{Number of actually used tools}} = \frac{|R_i \cap U_i|}{|U_i|}, U_i > 0$$

RQ2: Evaluation of MCP-Enhanced LLMs

Experiment 2: MCP Effectiveness over Multi-models

Model	Model Scale	Model Characteristics	Experimental Conditions
Qwen2.5-7B-Instruct-Turbo (small)	7B parameters	Lightweight model with limited internal knowledge	<ul style="list-style-type: none">• Baseline vs MCP• Identical task instructions and test dataset
Mistral-Small-24B-Instruct-2501 (medium)	24B parameters	Balanced reasoning and efficiency; mid-scale instruction-tuned model	
Qwen-plus (large)	>100B parameters	Strong intrinsic reasoning ability; extensive internal knowledge	

RQ2: Evaluation of MCP-Enhanced LLMs

Experiment 2: MCP Effectiveness over Multi-models

LLM (size)	Correctness without tool using	Correctness with MCP tool using	Tool Usage Frequency	Tool Usage Precision
Qwen-plus (large)	95%	100% (+5%) -	28% ↓↓	94% ↑↑
Mistral-Small-24B-Instruct-2501 (medium)	52%	60% (+8%) ↑	74% -	92% ↑
Qwen2.5-7B-Instruct-Turbo (small)	31%	47% (+16%) ↑↑	66% -	82% -

Correctness (Baseline v.s. MCP)

- **Larger models** have **higher correctness**
- **Smaller models** **benefit more** from MCP integration

Tool Usage **Frequency**

- **Large models** **use less** MCP
- **Medium models** use MCP **slightly more** than **small models**
- Overall, **smaller models** tend to **rely more** on MCP

Tool Usage **Precision**

- **Large models** use MCP **more accurately**
- Tool usage precision **decreases with smaller model size**

- Task performance highly correlated to model size
- MCP improves all models, compensating for the gap between small model and large models

Conclusion and Outlook

Conclusion and Outlook

Key Contributions

- Design and implementation of an MCP-based scientific LLM system
- Construction of an MCP-ready test dataset adapted from authoritative benchmarks
- Comprehensive evaluation across models of different scales
- Key insights:
 - MCP is most necessary in parameter calculation and scientific data lookup tasks
 - MCP significantly improves small model performance, but show limit improvements for large models

Outlook

- Expanding tool coverage
- Increasing data sources for each scientific domains
- Scaling up the dataset for more robust evaluation

Source

- [1] Rahman, Roussel, and Aashwin Ananda Mishra. 2025. "A Fragile Number Sense: Probing the Elemental Limits of Numerical Reasoning in LLMs." ArXiv.org. 2025. <https://arxiv.org/abs/2509.06332>.
- [2] Boye, Johan, and Birger Moell. 2025. "Large Language Models and Mathematical Reasoning Failures." ArXiv.org. 2025. <https://arxiv.org/abs/2502.11574>.
- [3] Zhang, L., Graf, E., et al. (2025). Mathematical Computation and Reasoning Errors by Large Language Models. Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con 2025).
- [4] Peters, U., & Chin-Yee, B. (2025). Generalization Bias in Large Language Model Summarization of Scientific Research. arXiv preprint arXiv:2504.00025.
- [5] Anthropic. 2024. "Model Context Protocol: An Open Standard for Tool-Augmented LLM Systems." Technical specification. <https://modelcontextprotocol.io>
- [6] Wang, Y., Ma, X., Zhang, G., Ni, Y., and Chandra, A. 2024. "MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark." ArXiv.org. 2024. <https://arxiv.org/abs/2406.01574>.
- [7] Chen, W., Yin, M., Ku, M., Lu, P., Wan, Y., Ma, X., Xu, J., Wang, X., and Xia, T. 2023. "TheoremQA: A Theorem-Driven Question Answering Dataset." ArXiv.org. 2023. <https://arxiv.org/abs/2305.12524>.
- [8] Welbl, J., Stenetorp, P., and Clarke, T. 2017. "Crowdsourcing Multiple Choice Science Questions." ArXiv.org. 2017. <https://arxiv.org/abs/1707.06209>.



University of Stuttgart
Institut of Industrial Automation
and Software Engineering

Thank you!



Xincheng Chen

e-mail st190250@stud.uni-stuttgart.de

phone +49 (0) 711 685-

fax +49 (0) 711 685-

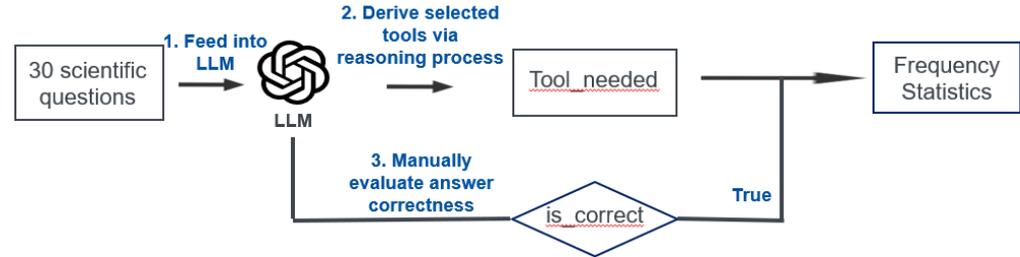
University of Stuttgart
Institut of Industrial Automation and Software Engineering
Pfaffenwaldring 47, 70550 Stuttgart, Germany



Frequency Analysis of Required Tool Types

Criteria

- Extract tool cues from the reasoning process
- Use only correct-answer traces as reliable evidence



Example from MMLU-Pro

Question: The metal beaker of a dew-point apparatus is gradually cooled from room temperature, 75°F. When the beaker temperature reaches 60°F, the moisture of the room air starts condensing on it. Assuming the room air to be at 14.7 psia, determine (a) the partial pressure of vapor, and (b) the parts by mass of vapor in the room air.

Reasoning Process (Excerpt): “Okay, let’s see. The problem is about a dew-point apparatus. ... First, part (a) asks for the partial pressure of vapor. ... So, I need to find the saturation pressure of water at 60°F. Wait, but the **units here are in Fahrenheit and psia**. I might **need a steam table or a formula** to calculate the saturation pressure at 60°F. Since I don’t have a steam table memorized, maybe I should use the **Antoine equation** or some approximation ...”

Identified Tool-Related Cues

- “need a steam table or a formula” → Physics database
- “Antoine equation “ → Web Search
- “units in °F and psia” → Calculator

Tool-Needed: { physics database, calculator, web search }

RQ2: Evaluation of MCP-Enhanced LLMs

System Overview and Integrated Tools

