



Carnegie Mellon University
Language Technologies Institute

Language Models & Tools

Zora (Zhiruo) Wang
Language Technologies Institute

LMs are powerful for text generation tasks.

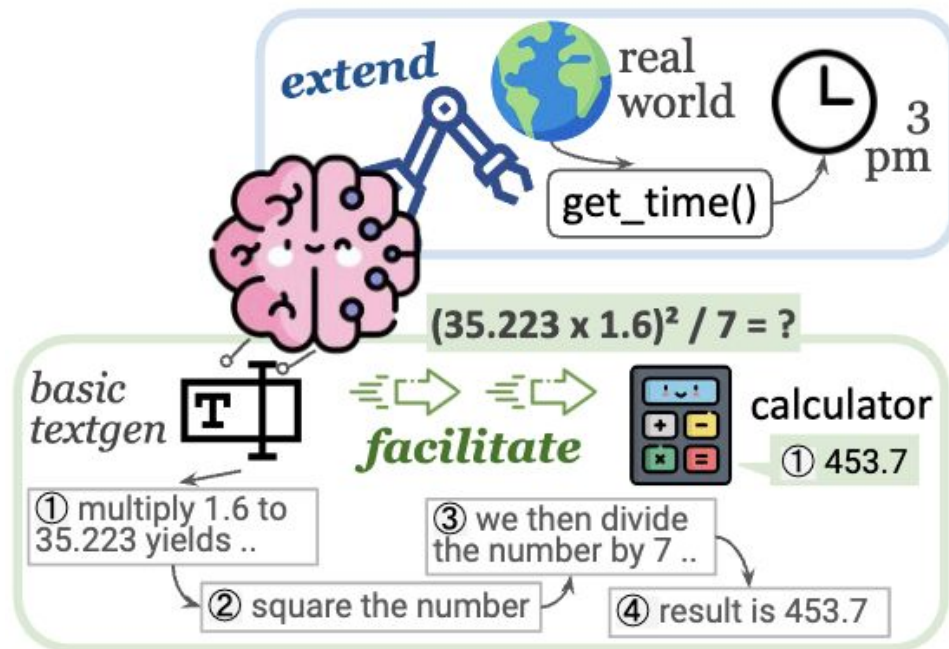
But ...

- Complex reasoning?

Struggle

- Access real-world information?

Fundamentally unable



Tools benefit language models a lot

- ToolFormer

The New England Journal of Medicine is a registered trademark of [QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society] the MMS.

Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

The name derives from "la tortuga", the Spanish word for [MT("tortuga") → turtle] turtle.

The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

ART: Automatic multi-step reasoning and tool-use for large language models

 **Bhargavi Paranjape** **TOOLLLM: FACILITATING LARGE LANGUAGE MODELS TO MASTER 16000+ REAL-WORLD APIS**

¹U

On the Tool Manipulation Capability of Open-source Large Language Models

¹, Lan Yan¹, Yaxi Lu¹, Yankai Lin^{3†}, Yuan Han¹ Dunshu Tian¹

Gorilla: Large Language Model Connected with Massive APIs

HuggingGPT: Solving AI Tasks with ChatGPT and its

Fried VOYAGER: An Open-Ended Embodied Agent with Large Language Models

TROVE: Inducing Verifiable and Efficient Toolboxes for Solving Programmatic Tasks

{syl,

Zhiruo Wang¹ Graham Neubig¹ Daniel Fried¹

dlekar^{1*}, andkumar^{1,2†}
ison
uthors

Tools benefit language models a lot

software

- ToolFormer

The New England Journal of Medicine is a registered trademark of [QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society] the MMS.

Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

The name derives from "la tortuga", the Spanish word for [MT("tortuga") → turtle] turtle.

The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

ART: Automatic multi-step reasoning and tool-use for large language models

Bhargavi Paranjape¹ Scott Lundberg² Sameer Singh³ Hannaneh Hajishirzi^{1,4} Luke Zettlemoyer^{1,5} Marco Tulio Ribeiro²

¹University of Washington, ²Microsoft Research, ³University of California, Irvine, ⁴Allen Institute of Artificial Intelligence, ⁵Meta AI

Open-source Large Language Models

Ruobing Xie¹, Ji Zhang¹, Mark Gerstein¹, Dehai Li¹, Zhenyu Liu¹, Maosen Sun^{1†}
¹Tsinghua University ²ModelBest Inc. ³Renmin University of China
⁴Yale University ⁵WeChat AI, Tencent Inc

Gorilla: Large Language Model Connected with Massive APIs

HuggingGPT: Solving AI Tasks with ChatGPT and its Friends

VOYAGER: An Open-Ended Embodied Agent with Large Language Models

TROVE: Inducing Verifiable and Efficient Toolboxes for Solving Programmatic Tasks

Yong Ge¹, Yixuan Li¹, Yuxuan Yao¹, Yuxin Chen¹, Yuxuan Wang¹, Dongsheng Li¹, Weimin Lu¹, Cheng Zhuang¹, Zhejiang University¹, MGuangzhi Wang^{1,2}, Yuqi Xie³, Yunfan Jiang^{4*}, Ajay Mandlekar^{1*}, Chaowei Xiao^{1,5}, Yuke Zhu^{1,7}, Linxiang Jun⁷, Fanyi Han¹, Anima Anandkumar^{1,2†}
¹NVIDIA, ²Caltech, ³UT Austin, ⁴Stanford, ⁵UW Madison
<https://github.com/mguangzhiwang> [†]Equal contribution [‡]Equal advising [✉]Corresponding authors
Zhiruo Wang¹ Graham Neubig¹ Daniel Fried¹

LANGUAGE
WORLD APIS

xi Lu¹, Yankai Lin^{3†},

Tools benefit language models a lot

- ToolFormer

The New England Journal of Medicine is a registered trademark of [QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society] the MMS.

Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

The name derives from "la tortuga", the Spanish word for [MT("tortuga") → turtle] turtle.

The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

ART: Automatic multi-step reasoning and tool-use for

APIs

Bhargavi Paranjape

1U

On the Open

Qiantong Xu, F



TOOLLLM: FACILITATING LARGE LANGUAGE MODELS TO MASTER 16000+ REAL-WORLD APIS

Yujia Qin^{1*}, Shihao Liang^{1*}, Yining Ye¹, Kunlun Zhu¹, Lan Yan¹, Yaxi Lu¹, Yankai Lin^{3†}, Xin Cong¹, Xiangru Tang⁴, Bill Qian⁴, Sihan Zhao¹, Lauren Hong¹, Runchu Tian¹, Ruobing Xie⁵, Jie Zhou⁵, Mark Gerstein⁴, Dahai Li^{2,6}, Zhiyuan Liu^{1†}, Maosong Sun^{1†}

¹Tsinghua University ²ModelBest Inc. ³Renmin University of China

⁴Yale University ⁵WeChat AI, Tencent Inc. ⁶Zhihu Inc.

yujiaqin16@gmail.com

Palo Alto, CA, USA

{qiantong.xu, jian.zhang}@sambanovsystems.com

Embodied Agent

with Large Language Models

TROVE: Inducing Verifiable and Efficient Toolboxes

for Solving Programmatic Tasks

Yongqiang Sun^{1,2}, Kaiqi Song^{2,3,4}, Xu Tan², Dongsheng Li¹, Weiming Lu^{1,4}, Feting Zhuang¹

Zhejiang University¹, Microsoft Research Asia², Yuqi Xie³, Yunfan Jiang^{4*}, Ajay Mandlekar^{1*}, {syl, luwm, yzhuang}@zju.edu.cn, {kaiqi.song, yuqi.xie, yunfan.jiang, ajay.mandlekar}@microsoft.com, Anhma Anandkumar^{1,2†}

¹NVIDIA, ²Caltech, ³UT Austin, ⁴Stanford, ⁵UW Madison

https://github.com/microsoft/trove Equal advising Corresponding authors

Zhiruo Wang, Graham Neubig, Daniel Fried

Tools benefit language models a lot

- ToolFormer

The New England Journal of Medicine is a registered trademark of [QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society] the MMS.

Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

The name derives from "la tortuga", the Spanish word for [MT("tortuga") → turtle] turtle.

The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

ART: Automatic multi-step reasoning and tool-use for large language models

Bhargavi Paranjape¹ Scott Lundberg² Sameer Singh³ Hannaneh Hajishirzi^{1,4}
Luke Zettlemoyer^{1,5} Marco Tulio Ribeiro²

¹University of Washington, ²Microsoft Research, ³University of California, Irvine,
⁴Allen Institute of Artificial Intelligence, ⁵Meta AI

Neural Models

Gorilla: Large Language Model Connected with Massive APIs

HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face

Joseph E. Gonzalez¹
arch

Yongliang Shen^{1,2,*}, Kaitao Song^{2,*†}, Xu Tan²,
Dongsheng Li², Weiming Lu^{1,†}, Yueting Zhuang^{1,†}
Zhejiang University¹, Microsoft Research Asia²
{syl, luwm, yzhuang}@zju.edu.cn, {kaitaosong, xuta, dongсли}@microsoft.com

<https://github.com/microsoft/JARVIS>

Ajay Mandlekar^{1*},
Anima Anandkumar^{1,2†}
UW Madison
pending authors

Tools benefit language models a lot

- ToolFormer

The New England Journal of Medicine is a registered trademark of [QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society] the MMS.

Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

The name derives from "la tortuga", the Spanish word for [MT("tortuga") → turtle] turtle.

The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

ART: Automatic multi-step reasoning and tool-use for large language models

Bhargavi Paranjape¹ Scott Lundberg² Sameer Singh³ Hannaneh Hajishirzi^{1,4} Luke Zettlemoyer^{1,5} Marco Tulio Ribeiro²

¹University of Washington, ²Microsoft Research, ³University of California, Irvine, ⁴Allen Institute of Artificial Intelligence, ⁵Meta AI

Open-source LLMs to Master 600+ Real-World APIs
Xin Cong¹, Xiao Lu², Tang⁴, Bill Qian⁴, Sihao Zhao¹, Lauren Hong¹, Runchu Tian¹, Ruobing Xie⁵, Ji Zhang⁵, Mark Gerstein⁴, Dehai Li³, Zhipu Liu^{1†}, Maorion Sun^{1†}

Gorilla: Large Language Model Connected with Massive APIs

HuggingGPT: Solving AI Tasks with ChatGPT and its Friends

VOYAGER: An Open-Ended Embodied Agent with Large Language Models

TROVE: Inducing Verifiable and Efficient Toolboxes for Solving Programmatic Tasks

Zhiruo Wang¹ Graham Neubig¹ Daniel Fried¹

llekar^{1*},
andkumar^{1,2†}
son
authors

Expert-crafted
functions

What Is A Tool Anyway?

- Tool Basics: definition & functionality
- Scenarios: what tools, what tasks, what methods
- Evaluation, empirical benefit, future directions

What Is A Tool Anyway?

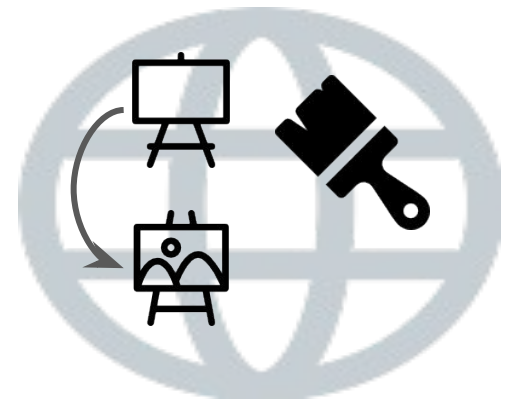
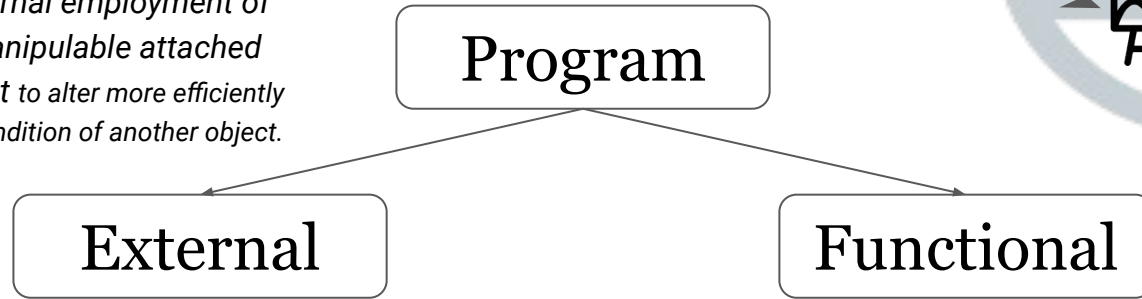
- **Tool Basics: definition & functionality**

- Scenarios: what tools, what tasks, what methods

- Evaluation, empirical benefit, future directions

Tool Basics: Definition

[1] *Animal tool*: the external employment of an unattached or manipulable attached environmental object to alter more efficiently the form, position, or condition of another object.



An LM-used tool is a **function** interface to a computer **program** that runs **external** to the LM, where the LM generates the function calls and input arguments in order to use the tool.

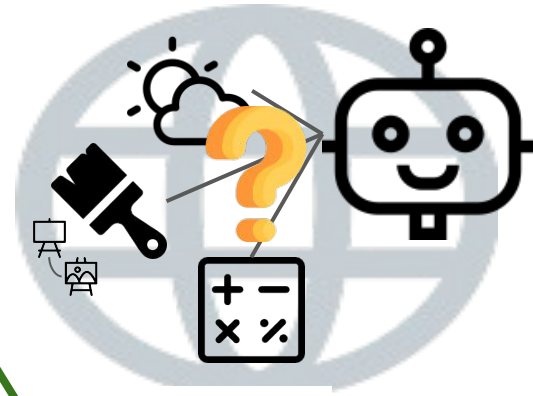
[1] Shumaker et al. *Animal tool behavior: the use and manufacture of tools by animals*. JHU Press, 2011.

Tool Basics: Functionality

 Perception: collect data from the env

 Action: exert actions, change env state

 Computation: general acts of computing



Tools

*Agents: anything that can be viewed as **perceiving** its environment through sensors and **acting** upon that environment through actuators^[1].*

[1] Russell, Stuart J., and Peter Norvig. *Artificial intelligence: a modern approach*. Pearson, 2016.

What Is A Tool Anyway?

- Tool Basics: definition & functionality

- **Scenarios: what methods, what tasks, what tools**

- Evaluation, empirical benefit, future directions

The Basic Tool Use Paradigm

Tool Use: switching between

- text-generation mode
- tool-execution mode

Tool Learning:

- inference-time prompting
- learning by training

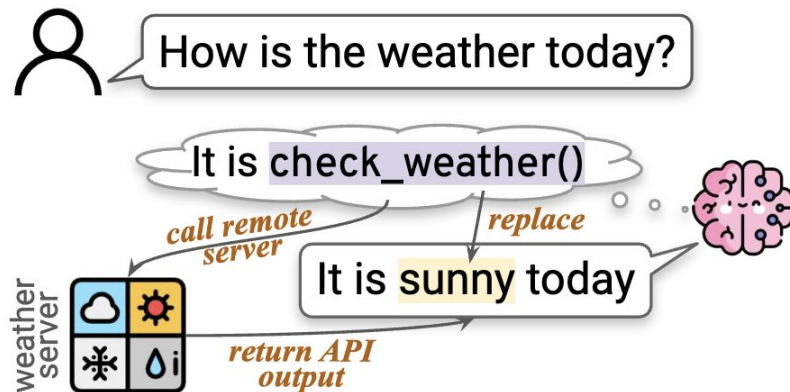


Figure 2: The basic tool use paradigm. LM calls `check_weather` tool by generating text tokens. This call triggers the server to execute the call and return the output `sunny`, using which the LM replaces the API call tokens in the response to the user.

Scenarios of LM Tool Using






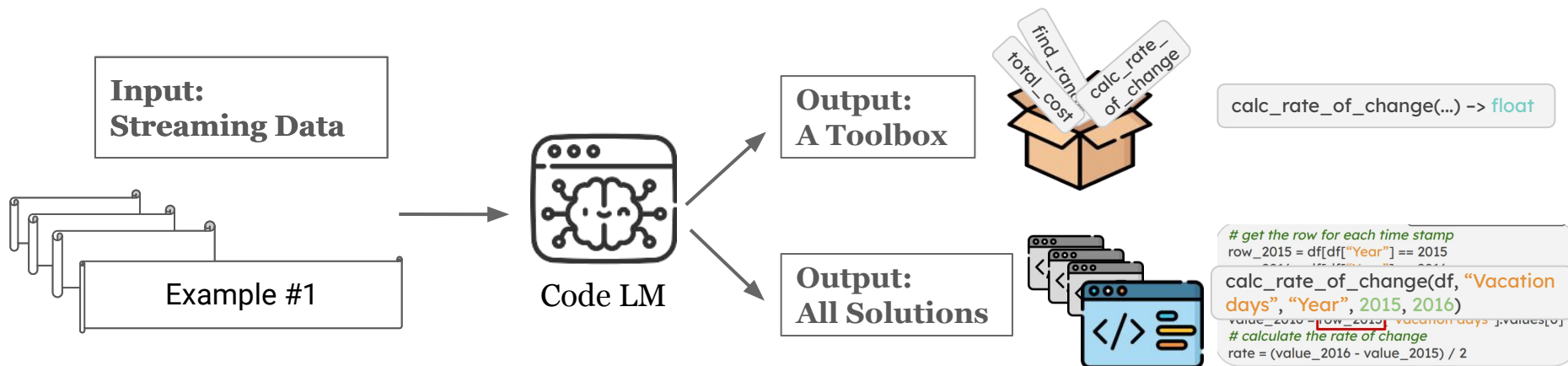
Category	Example Tools
 Knowledge access	<code>sql_executor(query: str) -> answer: any</code> <code>search_engine(query: str) -> document: str</code> <code>retriever(query: str) -> document: str</code>
 Computation activities	<code>calculator(formula: str) -> value: int float</code> <code>python_executor(program: str) -> result: any</code> <code>worksheet_get_row(row: list, index: int) -> None</code>
 Interaction w/ the world	<code>get_weather(city_name: str) -> weather: str</code> <code>get_location(city: str) -> location: str</code> <code>calendar_search(date: str) -> events: list</code> <code>email_checker(email: str) -> result: bool</code>
 Non-textual modalities	<code>cat_image(image_id: str) -> None</code> <code>spotify_get_song(song_name: str) -> None</code> <code>visual_qa(query: str, image: Image) -> answer: str</code>
 Special-skilled LMs	<code>QA(question: str) -> answer: str</code> <code>translation(text: str, language: str) -> text: str</code>

Table 1: Exemplar tools for each category.

What if tools are unavailable?

TROVE: Inducing Verifiable and Efficient Toolboxes for Solving Programmatic Tasks

Zhiruo Wang¹ Graham Neubig¹ Daniel Fried¹



How do TroVE make tools?

Pipeline

CREATE

IMPORT

SKIP

How can TroVE help?

Accuracy ↑

Complexity ↓

Verification ↑

Method	MATH _{algebra}		TabMWP		GQA	
	acc ↑	# lib ↓	acc ↑	# lib ↓	acc ↑	# lib ↓
<i>w/ additional supervision</i>						
LATM	0.30	-	0.09	-	0.29	
CRAFT	0.68	282	0.88	181	0.45	
<i>w/ additional rectification & iteration</i>						
Creator	0.65	875	0.81	4,595	0.34	
<i>w/o supervision, rectification, or iteration</i>						
TroVE	0.72	16	0.92	38	0.44	

Table 3. Comparing with existing methods using GPT-4. We report the baseline results as reported in Yuan et al. (2023). We do not report the *complexity* metric since none of these methods report it (our results in Table 2).

Method	prealg		TABLEQA			VISUAL GQA
	prealg	precal	TabMWP	WTQ	HiTab	
10% more accurate	Accuracy ↑		Time (s) ↓			31-43% faster
	avg	std	avg	std		
	0.77	0.109	25.5			
	0.88	0.024	30.7			
	0.87	0.057	17.5			

Table 5. Human accuracy and time in verifying model-produced solutions with three methods experimented.

1 MATH, TABLEQA, and VISUAL tasks.





















What Is A Tool Anyway?

- Tool Basics: definition & functionality

- Scenarios: what tools, what tasks, what methods

- **Evaluation, empirical benefit, future directions**

How to evaluate tool use?

	Benchmark	Tool Source	Example Curation	Domain (§4.1)	Executable
●	ToolBench	existing dataset	adopted, human annotated	 	✓
	ToolBench	RapidAPI	model synthesized	 	✓
	ToolQA	existing dataset	model synthesized	 	✓
	ToolAlpaca	PublicAPIs	model synthesized	   	x
	API-Bank	PublicAPIs	human annotated	 	✓
	MetaTool	OpenAI Plugins	model synthesized	  	x
	Gorilla	HF, Torch, TF	model synthesized		x
	HuggingGPT	HF	human annotated		x*
	Task Bench	HF, PublicAPIs	model synthesized	  	x

- Naturalness
- Executability

- reproducible testing
- Safe usage

ools

Trade-offs in tool usage: Computation Cost 💰

What tasks benefit the most from tools?

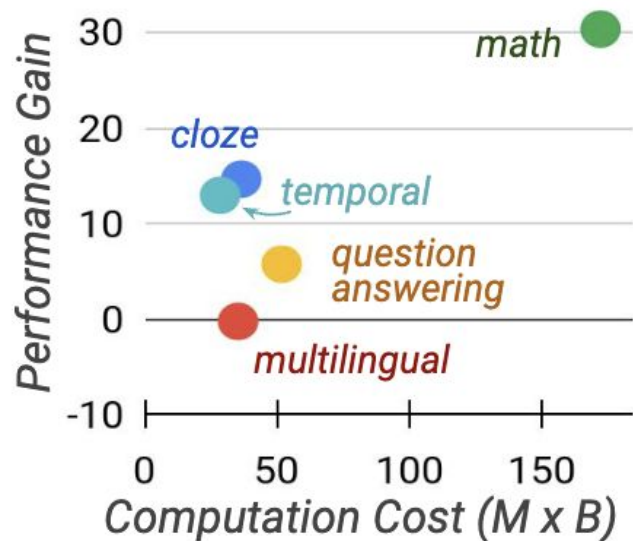


Figure 5: Compute & performance gain with ToolFormer.

What methods are efficient in tooling?

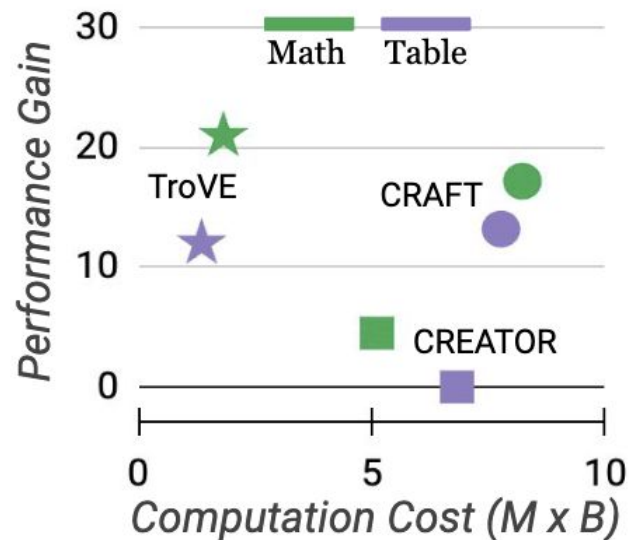


Figure 6: Comparing different tool-making methods.

In Summary

- Tool Basics: definition & functionality
- Scenarios: what tools, what tasks, what methods
- Evaluation, empirical benefit, future directions

Questions?