Compound AI Systems

Fatemeh Rahimi - March 2024

Hello I’m ...

Fatemeh

Senior NLP Scientist at Pythonic AI

- Train/Finetune Language Models
- Build LLM based workflows
- Build Agents

01

Recent Large Language Models

OpenAI
ChatGPT

Introducing GPT-4o

GPT-4o is our newest flagship model that provides GPT-4-level intelligence but is much faster and improves on its capabilities across text, voice, and vision.

Google The Keyword

Introducing Gemini 2.0: our new AI model for the agentic era

Dec 11, 2024 · 10 min read

Read AI-generated summary

Share

Claude 3.7 Sonnet and Claude Code

Introducing Claude 3.7 Sonnet, our most intelligent model yet and the first hybrid reasoning model. We're also launching Claude Code, an agentic tool for coding.

PaLM 2

A next generation language model with improved multilingual, reasoning and coding capabilities.

Alibaba xAI NVIDIA Meta MISTRAL AI_

OpenAI Google ANTHROP\C cohere AI21labs

OpenAI
ChatGPT

Introducing GPT-4o

GPT-4o is our newest flagship model that provides GPT-4-level intelligence but is much faster and improves on its capabilities across text, voice, and vision.

Google The Keyword

Introducing Gemini 2.0: our new AI model for the agentic era

Dec 11, 2024 · 10 min read

Read AI-generated summary

Share

Claude 3.7 Sonnet and Claude Code

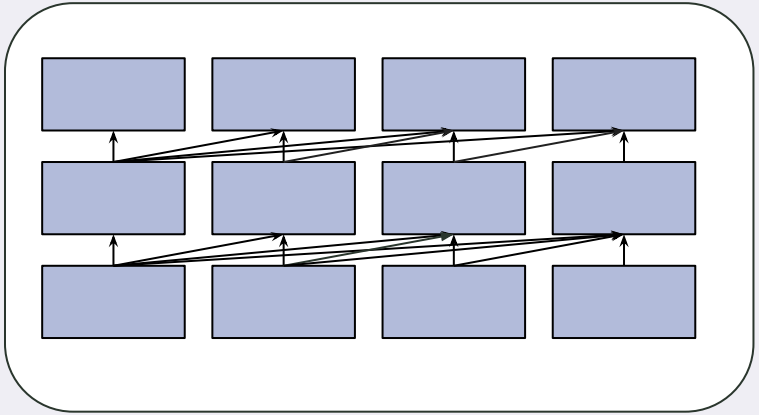
Introducing Claude 3.7 Sonnet, our most intelligent model yet and the first hybrid reasoning model. We're also launching Claude Code, an agentic tool for coding.

PaLM 2

A next generation language model with improved multilingual, reasoning and coding capabilities.

Models? Or Systems?

A model

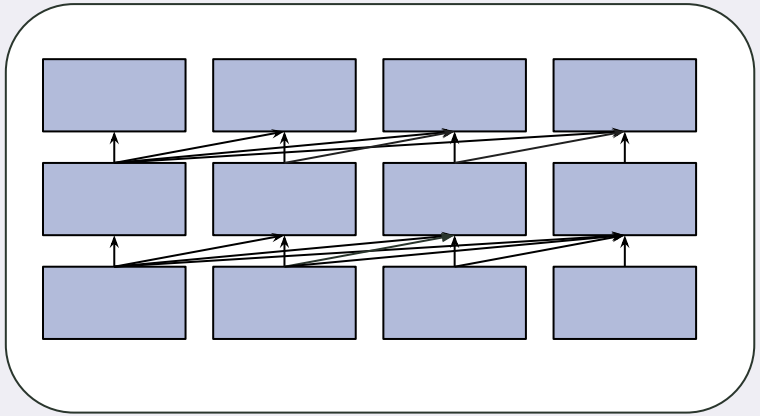


Zip File of the internet

Minimal System

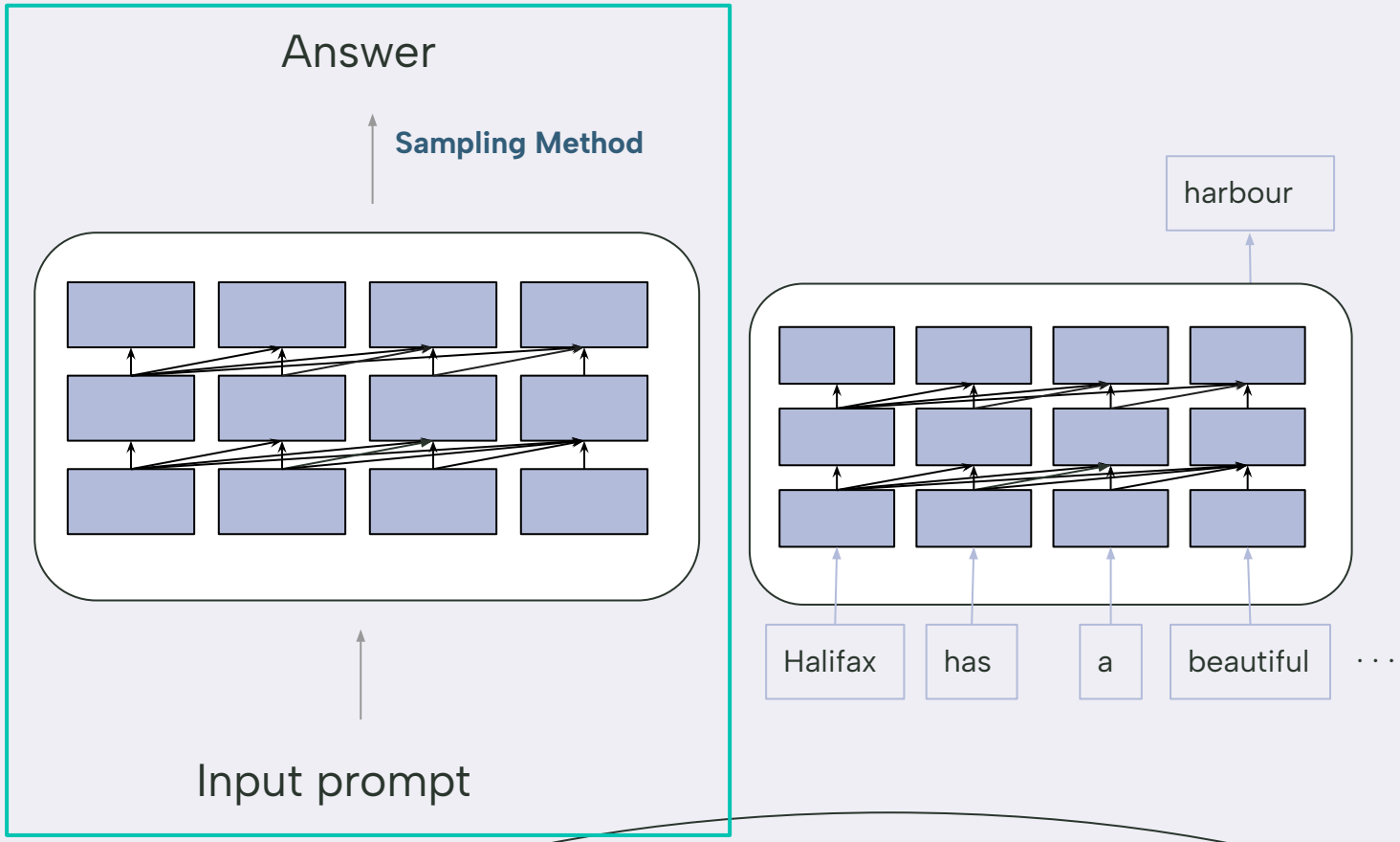
Answer

Sampling Method



Input prompt

Minimal System

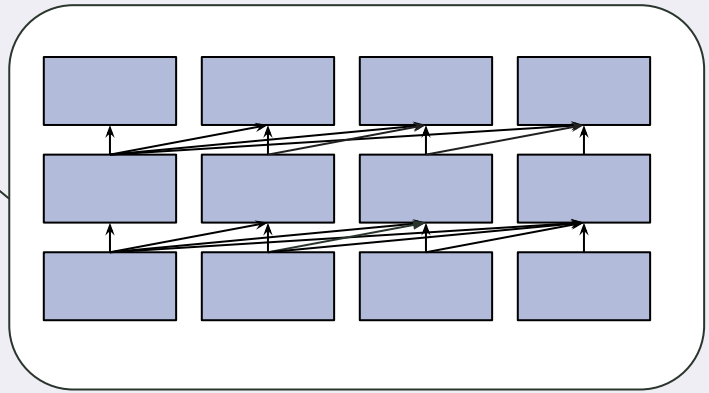


Recommended video: Transformers (how LLMs work) explained visually | DL5 by 3Blue1Brown

Minimal System

Answer

Sampling Method



Input prompt

Minimal System

Halifax has a beautiful

...harbour! 😄

Halifax has a beautiful harbour that's one of the largest natural harbours in the world.

Minimal System

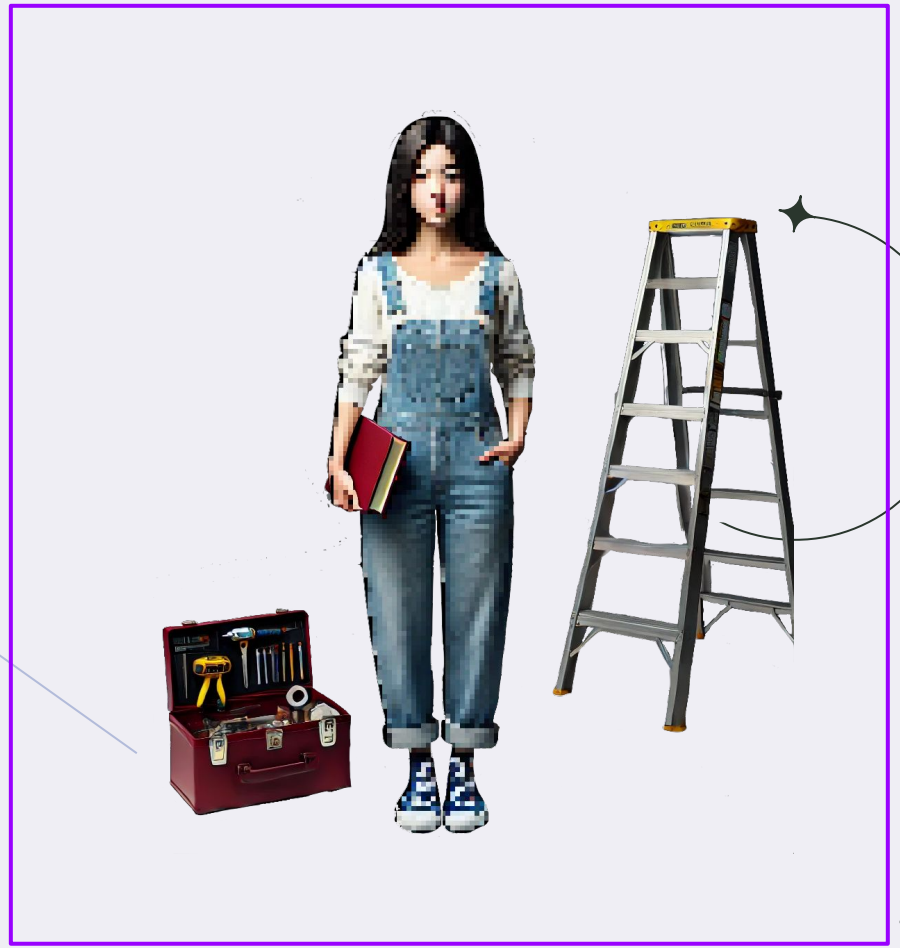
Halifax has a beautiful

...harbour! 😄

Halifax has a beautiful harbour that's one of the largest natural harbours in the world.

Advanced System

Advanced System



02

Why Systems are important?

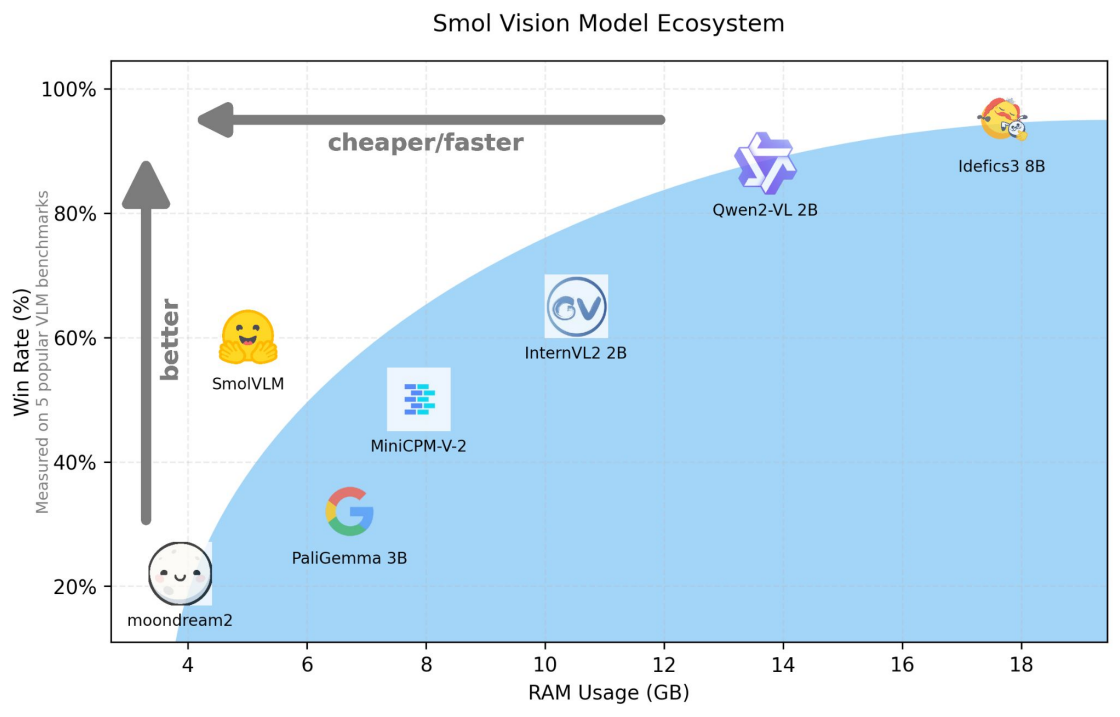
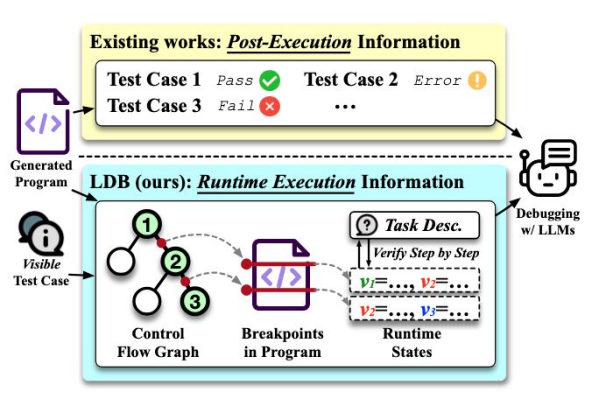


Image source: SmolVLM - small yet mighty Vision Language Model by Huggingface

Some tasks are easier to improve via System Design

Debug like a Human: A Large Language Model Debugger via Verifying Runtime Execution Step by Step

Li Zhong Zilong Wang[†] Jingbo Shang[†]
University of California, San Diego
{lizhong, zlwang, jshang}@ucsd.edu



Model (# Param.)	Debugger	Dataset					
		HumanEval		TransCoder		MBPP	
		Acc. ↑	Δ ↑	Acc. ↑	Δ ↑	Acc. ↑	Δ ↑
GPT-3.5 (≥175B[†])	Baseline (w/o debugger)	73.8		82.3		67.6	
	SD (+Expl.) (Chen et al., 2023c)	81.1	+7.3	85.9	+3.6	74.4	+6.8
	SD (+Trace) (Chen et al., 2023c)	80.5	+6.7	86.1	+3.8	72.6	+5.0
	LDB (ours)	**82.9**	**+9.1**	**87.7**	**+5.4**	**76.0**	**+8.4**
CodeLlama (34B)	Baseline (w/o debugger)	49.4		69.8		51.2	
	SD (+Expl.) (Chen et al., 2023c)	53.0	+3.6	79.4	+9.6	55.6	+4.4
	SD (+Trace) (Chen et al., 2023c)	54.3	+4.9	76.4	+6.6	57.2	+6.0
	LDB (ours)	**55.5**	**+6.1**	**79.6**	**+9.8**	**57.4**	**+6.2**
StarCoder (15B)	Baseline (w/o debugger)	39.0		61.8		51.6	
	SD (+Expl.) (Chen et al., 2023c)	38.4	-0.6	68.9	+7.1	54.4	+2.8
	SD (+Trace) (Chen et al., 2023c)	39.0	+0.0	65.7	+3.9	54.8	+3.2
	LDB (ours)	**39.6**	**+0.6**	**69.8**	**+8.0**	**55.4**	**+3.8**

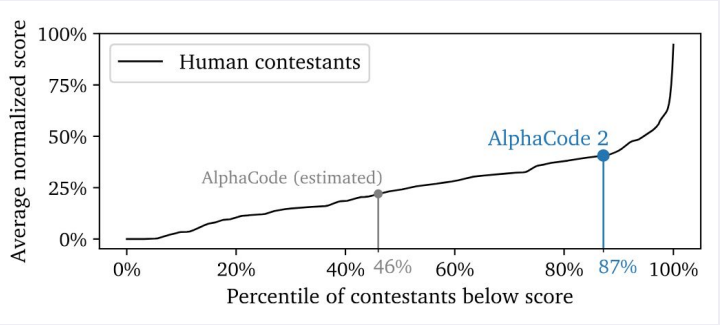
Table 1: Results of LDB and Self-Debugging (Chen et al., 2023c) (denoted as SD) on HumanEval, TransCoder, and MBPP with GPT-3.5, CodeLlama, and StarCoder. Accuracy is calculated based on Pass@1. The improvement (denoted as Δ) is measured against the baseline (w/o debugger). [†] We assume the parameter number in GPT-3.5 is larger than that of GPT-3 (175B).

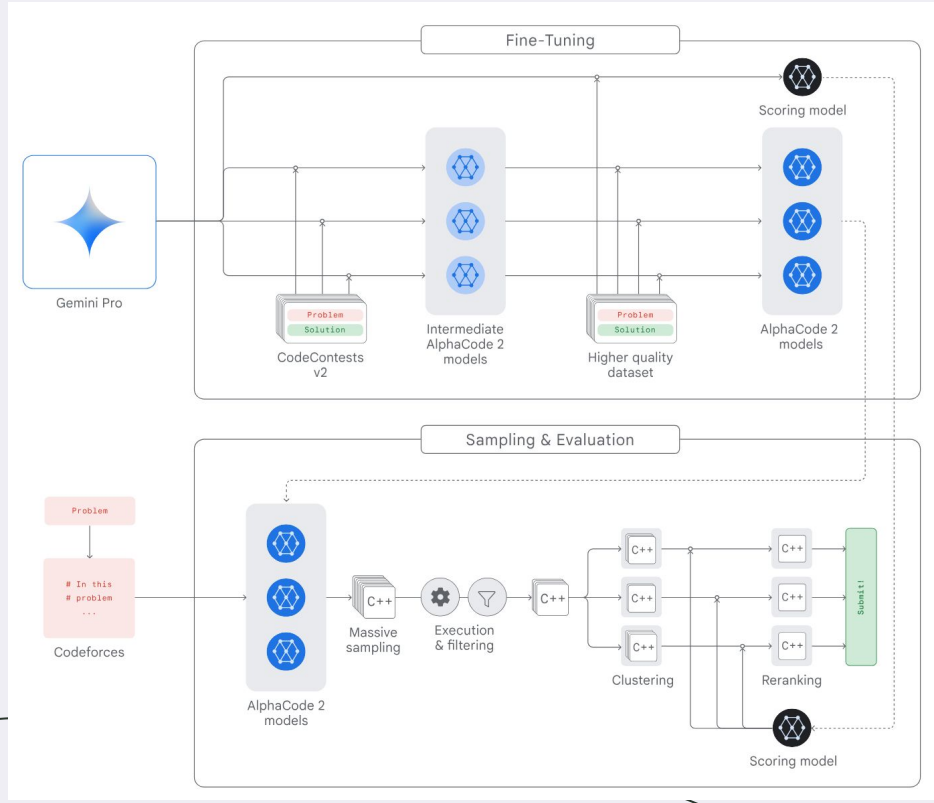
9.8% improvement

Some tasks are easier to improve via System Design

Google DeepMind

AlphaCode 2 Technical Report

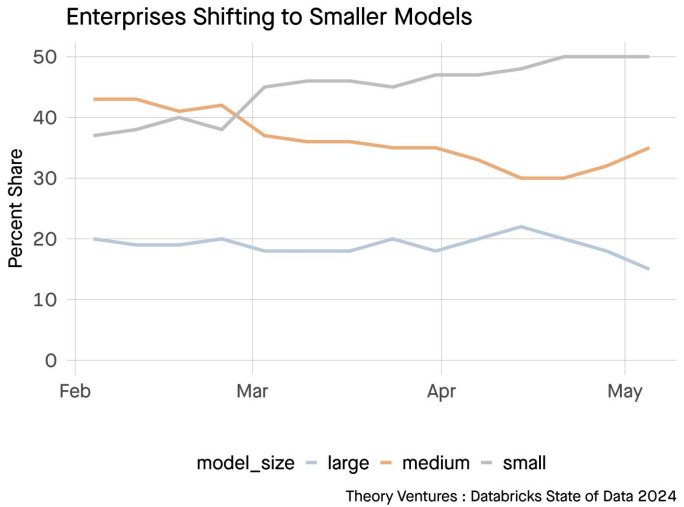




Some tasks are easier to improve via System Design

Small but Mighty AI

77% of enterprise AI usage are using models that are small models, less than 13b parameters.



Source: Small but Mighty AI by Tomasz Tunguz

Systems can be dynamic

LLM –> Limited to the knowledge during training

Systems can be dynamic

LLM -> Limited to the knowledge during training

System -> Access more information with Search and Retrieval

Why was my last bill higher than usual?

Your bill for March increased by $20 due to international calling charges on March 5 and 7. You can view the itemized list [here].

Safety and Regulations

Neural network models –> Hard to control during training

Preference tuning is a valid and well-researched approach, but it nearly impossible to guarantee the results.

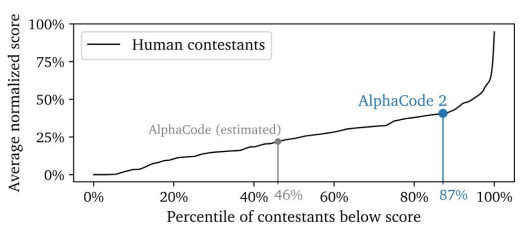
AI system –> can help control safety guidelines more tightly

by filtering model outputs

Read more about Preference tuning: Preference Tuning LLMs with Direct Preference Optimization Methods by Huggingface - Jan, 2024

Why Systems are important?

01 Some tasks are easier to improve via system design.



02 Systems can be dynamic.

03 Improving control and trust is easier with systems.

04 Performance goals vary widely.

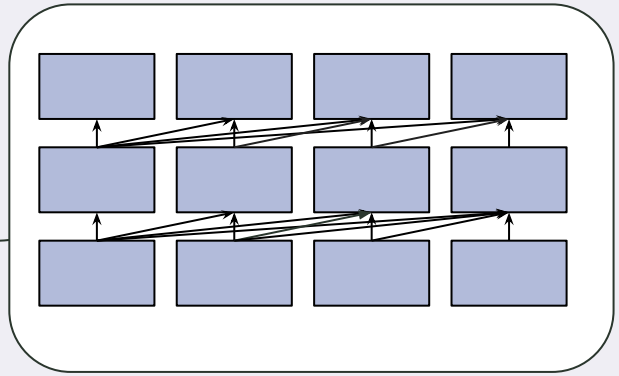
03

AI System Components

Minimal System

Answer

Sampling Method



Input prompt

Minimal System

Halifax has a beautiful

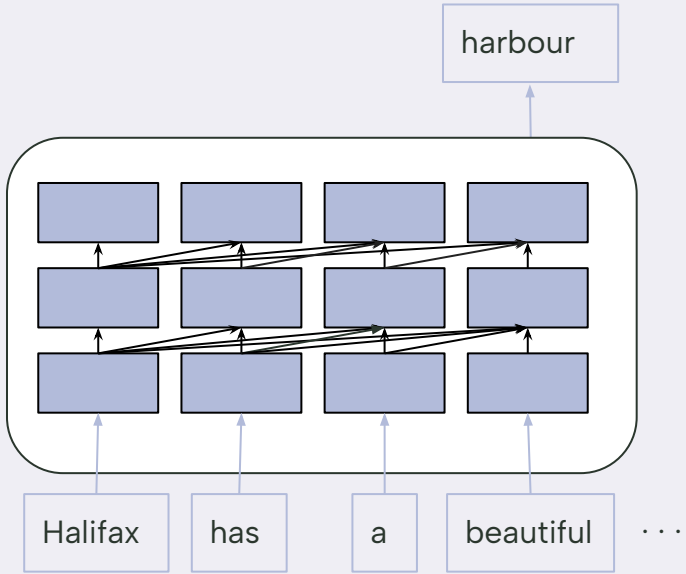
...harbour! 😄

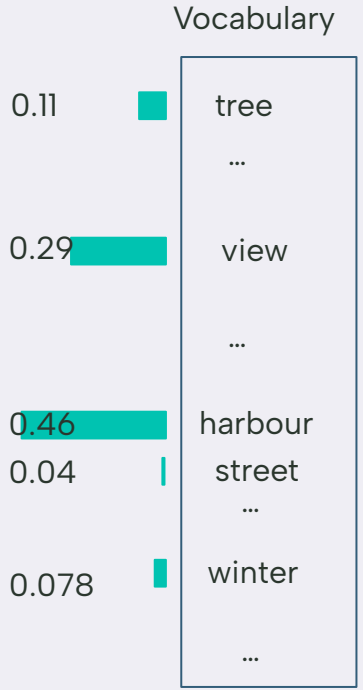
Halifax has a beautiful harbour that's one of the largest natural harbours in the world.

Minimal Systems

Sampling Methods

- Greedy
- Top – P
- Beam Search
- Token Diversity
- Valid json





Other constraint decoding method

Back to the Future: Unsupervised Backprop-based Decoding for Counterfactual and Abductive Commonsense Reasoning

Lianhui Qin[†‡] **Vered Shwartz** [†‡] **Peter West** [†‡] **Chandra Bhagavatula**[‡]
Jena D. Hwang [‡] **Ronan Le Bras** [‡] **Antoine Bosselut** [♮‡] **Yejin Choi**[†‡]
[†]Paul G. Allen School of Computer Science & Engineering, University of Washington
[‡]Allen Institute for Artificial Intelligence [♮]Stanford University
{lianhuiq, pawest, yejin}@cs.washington.edu
{vered, chandrab, jenah, ronanlb}@allenai.org

NEUROLOGIC DECODING: (Un)supervised Neural Text Generation with Predicate Logic Constraints

Ximing Lu[†‡] **Peter West**[†‡] **Rowan Zellers**[†‡]
Ronan Le Bras[‡] **Chandra Bhagavatula**[‡] **Yejin Choi**[†‡]
[†]Paul G. Allen School of Computer Science & Engineering, University of Washington
[‡]Allen Institute for Artificial Intelligence
{lux32, pawest, rowanz, yejin}@cs.washington.edu
{ronanlb, chandrab}@allenai.org

Grammar-Constrained Decoding for Structured NLP Tasks without Finetuning

Saibo Geng,[◇] **Martin Josifoski,**[◇] **Maxime Peyrard,**[*♣] **Robert West**[◇]
[◇]EPFL [♣]Université Grenoble Alpes, CNRS, Grenoble INP, LIG
{saibo.geng, martin.josifoski, robert.west}@epfl.ch, maxime.peyrard@univ-grenoble-alpes.fr

SynCode: LLM Generation with Grammar Augmentation

Shubham Ugare
University of Illinois
Tarun Suresh
University of Illinois
Hangoo Kang
University of Illinois
Sasa Misailovic
University of Illinois
Gagandeep Singh
University of Illinois

Table 6: Overview of various constrained decoding methods

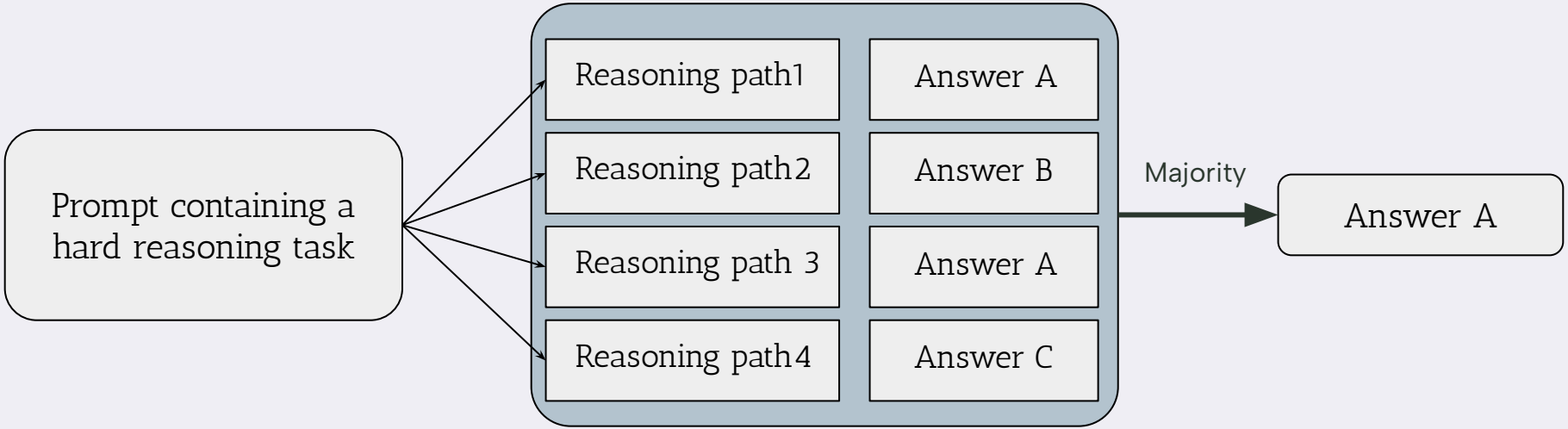
	Regex	CFG	Precomputed	GPL	Max CFG	Input format
LMQL (Beurer-Kellner et al., 2023)	✓	✗	✗	✗	50-100	LMQL DSL
GUIDANCE (Lundberg et al., 2023)	✓	✓	✗	✗	50-100	Python DSL
OUTLINES (Willard and Louf, 2023)	✓	✓	✓	✗	50-100	Lark EBNF
PICARD (Scholak et al., 2021)	✓	✓	✗	✗	50-100	Haskell
SYNCHROMESH (Poesia et al., 2022)	✓	✓	✗	✗	‡	ANTLR
LLAMA.CPP (Gerganov and et. al., 2024)	✓	✓	✗	✗	50-100	GBNF DSL
GCD (Geng et al., 2023)	✓	✓	✗	✗	50-100	GF
DOMINO (Beurer-Kellner et al., 2024)	✓	✓	✓	✗	50-100	GBNF DSL
SYNCODE (ours)	✓	✓	✓	✓	500+	Lark EBNF

† Implementation issues ‡ Synchromesh is closed-source and the information about DSL grammars is unavailable
GF: Grammatical Framework, GBNF is a DSL defined by LLAMA.CPP

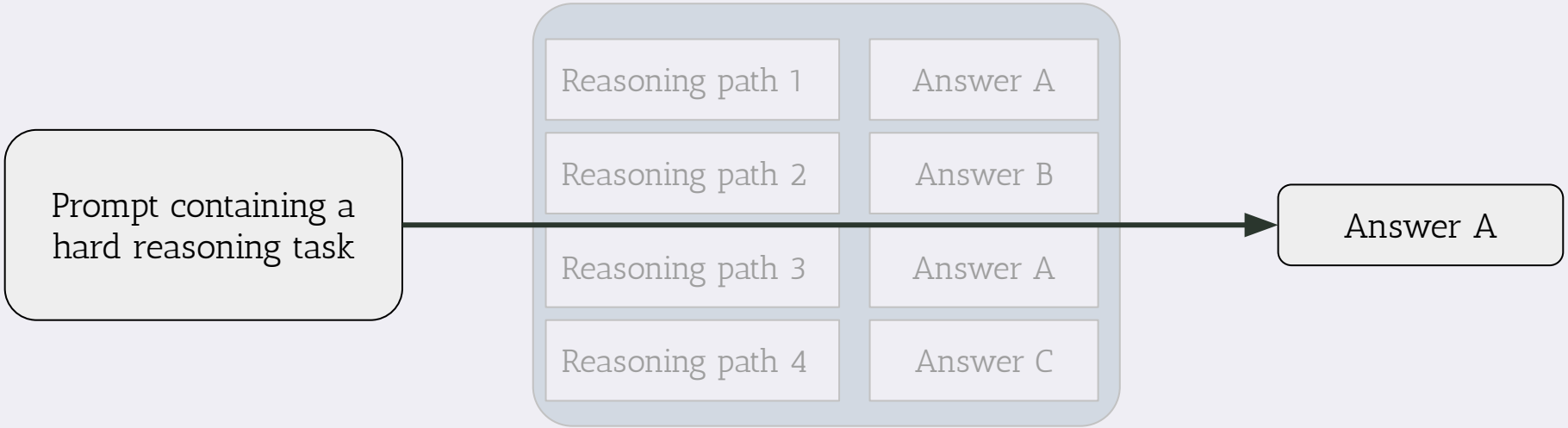
Adaptive Decoding via Latent Preference Optimization

Shehzaad Dhuliawala [1 2] **Ilia Kulikov** [1] **Ping Yu** [1] **Asli Celikyilmaz** [1]
Jason Weston [1] **Sainbayar Sukhbaatar** [1] **Jack Lanchantin** [1]

Mejority completion strategies



Mejority completion strategies



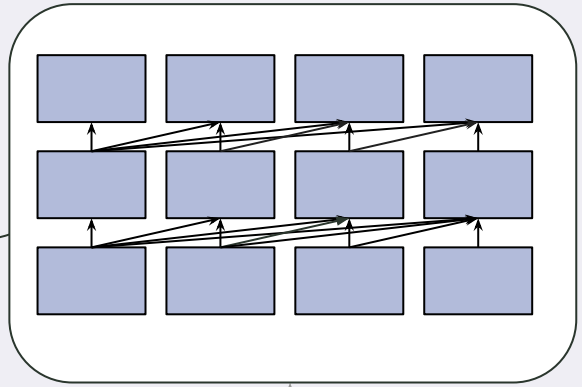
No “One True” sampling method

Your Choice will be highly impacting your results

Minimal System

Answer

Sampling Method



Input prompt

Minimal System

Halifax has a beautiful

...harbour! 😄

Halifax has a beautiful harbour that's one of the largest natural harbours in the world.

Prompting : The heart of modern AI System design

Prompting a new model is wild. It's like: add an example — works better. Add another example — still good. Add one more — boom, the entire essay is about cucumbers for no reason.

These new models are so sensitive, I swear I spent more time emotionally validating my prompt than writing it. 'You're doing great, model. You're so smart. Now please… summarize this article?

LLM generated Jokes :)

QUANTIFYING LANGUAGE MODELS' SENSITIVITY TO SPURIOUS FEATURES IN PROMPT DESIGN *or:* *How I learned to start worrying about prompt formatting*

Melanie Sclar[1] **Yejin Choi**[1,2] **Yulia Tsvetkov**[1] **Alane Suhr**[3]
[1]Paul G. Allen School of Computer Science & Engineering, University of Washington
[2]Allen Institute for Artificial Intelligence [3]University of California, Berkeley
msclar@cs.washington.edu

Table 2: Examples of atomic changes' impact on accuracy using probability ranking (prefix matching shown in Table 4). {} represents a text field; p_2 yields higher accuracy than p_1 for all tasks.

Task Id	Prompt Format 1 (p_1)	Prompt Format 2 (p_2)	Acc p_1	Acc p_2	Diff.
task280	`passage:{}\n answer:{}`	`passage {}\n answer {}`	0.043	0.826	0.783
task317	`Passage::{} Answer::{}`	`Passage:: {} Answer:: {}`	0.076	0.638	0.562
task190	`Sentence[I]- {}Sentence[II]- {} -- Answer\t{}`	`Sentence[A]- {}Sentence[B]- {} -- Answer\t{}`	0.360	0.614	0.254
task904	`input:: {} \n output:: {}`	`input::{} \n output::{}`	0.418	0.616	0.198
task320	`target - {} \n{} \nanswer - {}`	`target - {}; \n{}; \nanswer - {}`	0.361	0.476	0.115
task322	`COMMENT: {} ANSWER: {}`	`comment: {} answer: {}`	0.614	0.714	0.100
task279	`Passage : {}. Answer : {}`	`PASSAGE : {}. ANSWER : {}`	0.372	0.441	0.069

small, often semantically irrelevant changes in prompts -> dramatically change the outputs

Code Generation on HumanEval

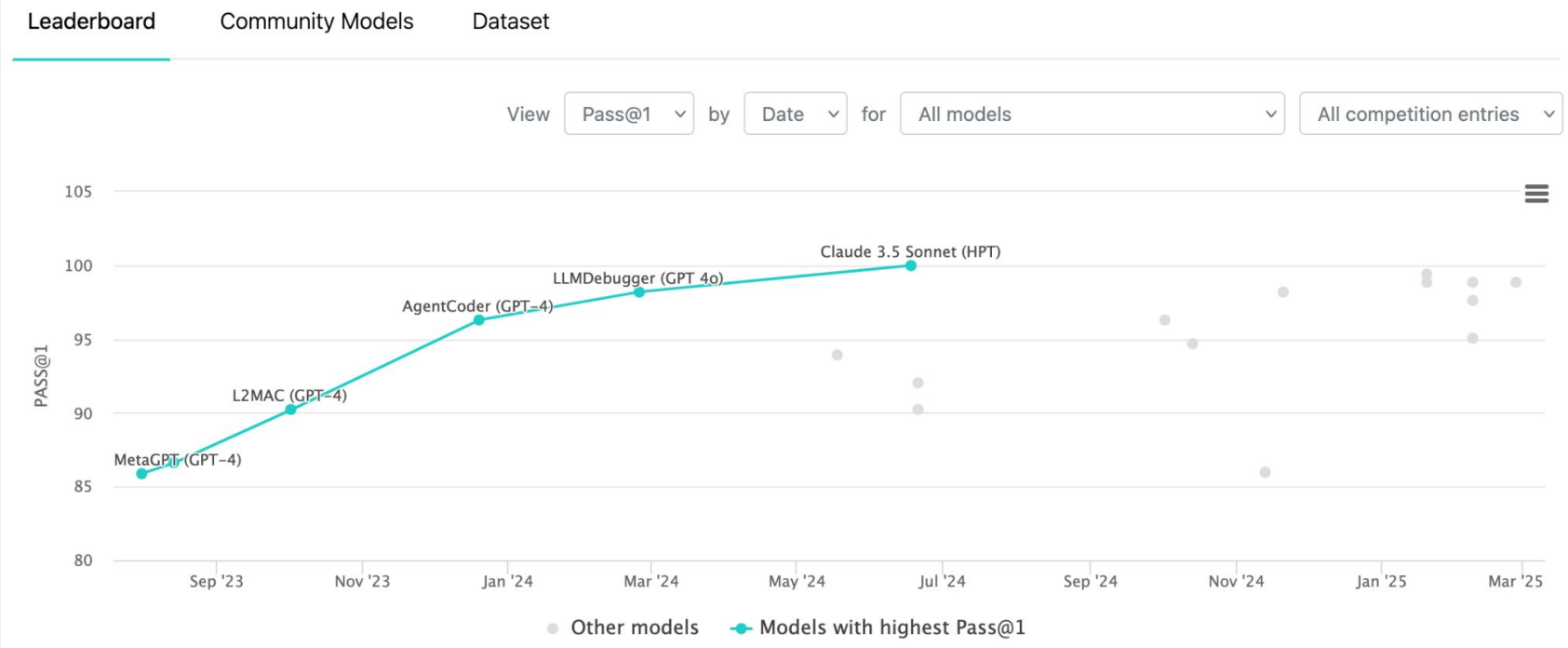
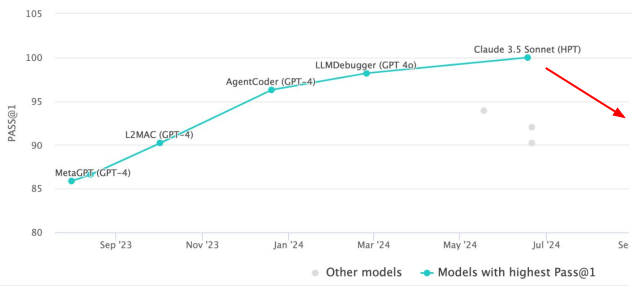


Image source: Paper with code, screenshot taken March, 2025

Code Generation on HumanEval

Leaderboard | Community Models | Dataset

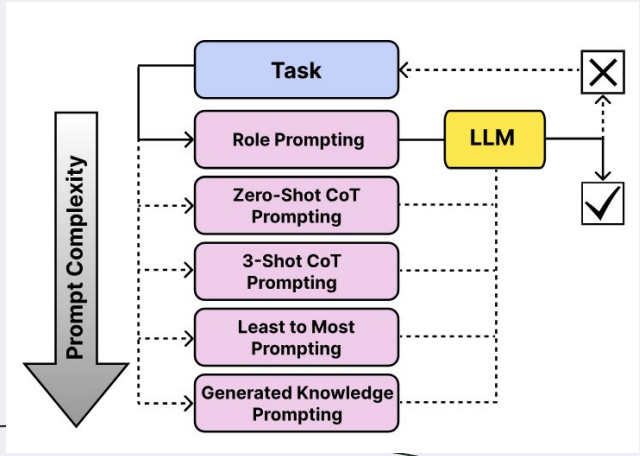
View Pass@1 by Date for All models All competition entries



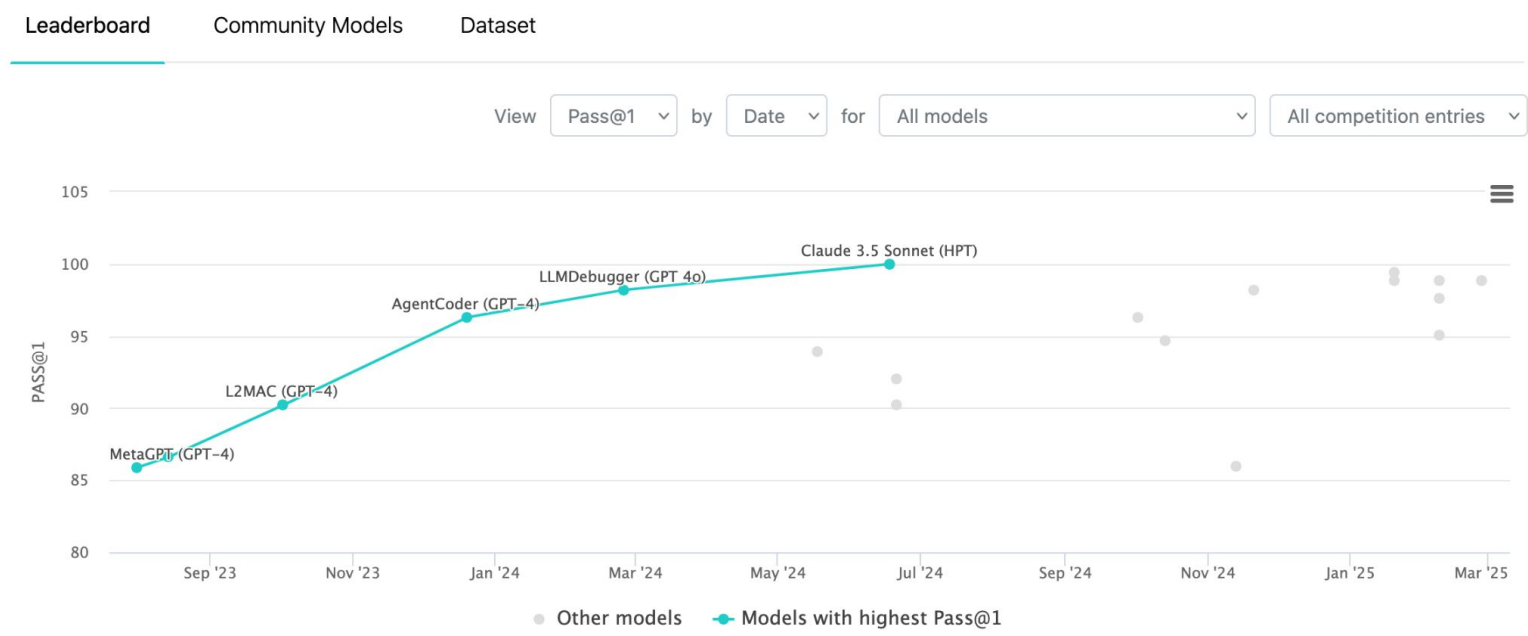
Hierarchical Prompting Taxonomy: A Universal Evaluation Framework for Large Language Models Aligned with Human Cognitive Principles

Devichand Budagam[1], **Ashutosh Kumar**[2], **Mahsa Khoshnoodi**[3], **Sankalp KJ**[4], **Vinija Jain**[5, 7 *], **Aman Chadha**[6, 7 †]

[1]Indian Institute of Technology Kharagpur, India [2]Rochester Institute of Technology, USA
[3]Researcher, Fatima Fellowship [4]AI Institute, University of South Carolina, USA
[5]Meta, USA [6]Amazon GenAI, USA [7]Stanford University, USA



Code Generation on HumanEval



What is the optimal prompt/model combination?

EchoPrompt: Instructing the Model to Rephrase Queries for Improved In-context Learning

Rajasekhar Reddy Mekala* rmekala@uci.edu
Yasaman Razeghi * yrazeghi@uci.edu
Sameer Singh sameer@uci.edu

EchoPrompt?	Stage-1 Prompt	GSM8K	SVAMP	MultiArith	SingleOp
Zero-shot					
✗	-	16.4	66.8	31.0	91.6
✓	Let's repeat the question. "	**20.7**(+4.3)	**74.7**(+7.9)	48.5(+17.5)	91.8(+0.2)
✓	Let's reiterate the question. "	19.7(+3.3)	73.4(+6.6)	**51.0**(+20.0)	93.0(+1.4)
✓	Let's restate the question. "	19.2(+2.8)	74.6(+7.8)	47.7(+16.7)	89.6(−2.0)
✓	Let's summarize the question. "	20.6(+4.2)	73.2(+6.4)	48.8(+17.8)	**93.7**(+2.1)
Zero-shot-CoT					
✗	Let's think step by step.	49.3	66.5	76.0	82.9
✓	Let's repeat the question and also think step by step.	44.6(−4.7)	**74.7**(+8.2)	70.9(−5.1)	92.3(+9.4)
✓	Let's reiterate the question and also think step by step.	**51.1**(+1.8)	73.9(+7.4)	78.7(+2.7)	**92.4**(+9.5)
✓	Let's repeat the question and also think step by step. "	42.0(−7.3)	60.4(−6.1)	78.1(+2.1)	88.3(+5.4)
✓	Let's restate the question and also think step by step.	47.0(−2.3)	73.9(+7.4)	**79.3**(+3.3)	90.2(+7.3)
✓	Let's summarize the question and also think step by step.	49.9(+0.6)	74.2(+7.7)	75.8(−0.2)	90.9(+8.0)

Code-davinci-002 model

Center for Research on Foundation Models HELM Image2Struct

Model	Mean win rate
GPT-4o (2024-08-06)	**0.947**
GPT-4o (2024-05-13)	0.807
Gemini 1.5 Pro (002)	0.789
Claude 3.5 Sonnet (20240620)	0.737
Gemini 1.5 Flash (002)	0.711
GPT-4o mini (2024-07-18)	0.702
Gemini 1.5 Pro (0409 preview)	0.667
Claude 3 Opus (20240229)	0.623
Claude 3.5 Sonnet (20241022)	0.605
Claude 3 Sonnet (20240229)	0.491
Gemini 1.0 Pro Vision	0.474
GPT-4V (1106 preview)	0.456

🏆 Chatbot Arena LLM Leaderboard: Community-driven Evaluation for Best LLM and AI chatbots

Rank* (UB)	Rank (StyleC	Model	Arena Score
3	2	ChatGPT-4o-latest (2025-01-29)	1374
6	4	DeepSeek-R1	1360
6	11	Gemini-2.0-Flash-001	1355
6	3	o1-2024-12-17	1351
8	11	Gemma-3-27B-it	1341
9	11	Qwen2.5-Max	1340
9	7	o1-preview	1335

Let's think deeper?

Is this what we want to be evaluating?

Image source: CHatbot Arena Leaderboard

04

Developing compound AI System in Real world

Developing AI Systems

Software Engineers + Scientist

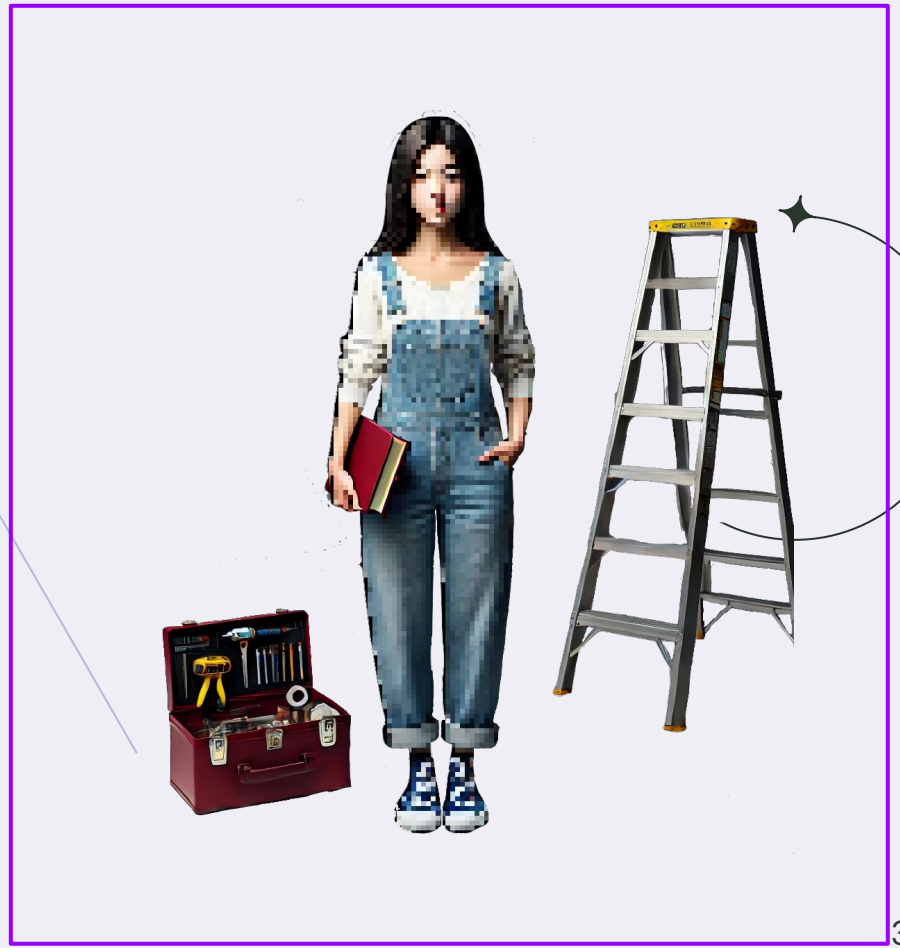
- Modular System Design
- Data Driven Optimization
- Generic Architecture

Avoid

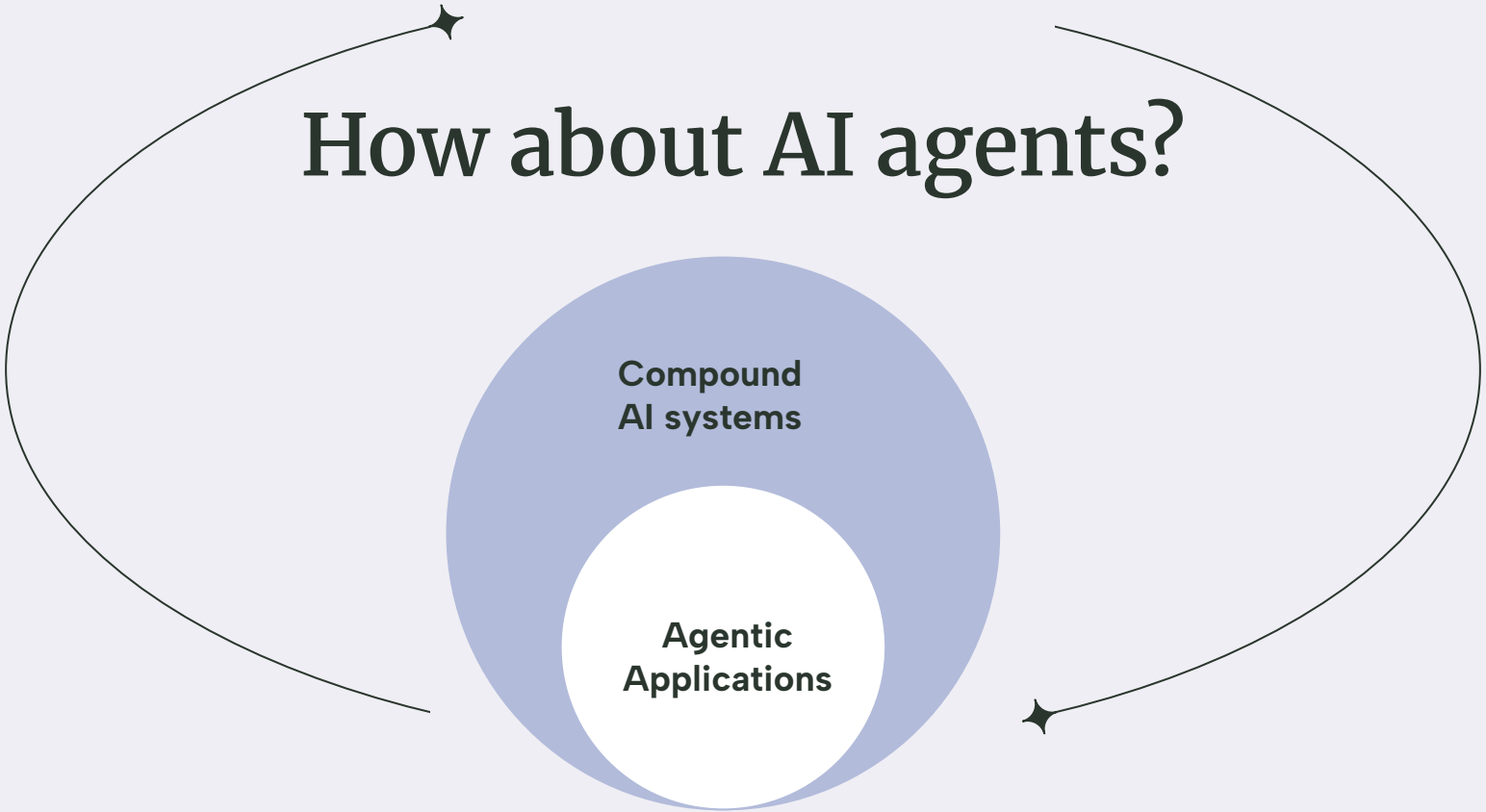
- Manul prompt adjustments
- Prompt templates
- Model dependence

Advanced System

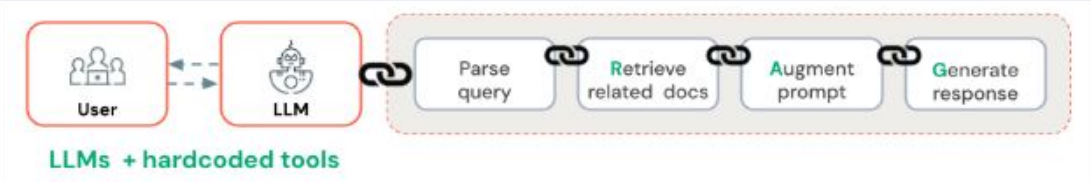
- LLM
- Web Browser plugin
- Code Interpreters
- Image generators
- Tool Execution
- Inference time prompting
- Verifications and filtering
- Clustering
- Language translation



How about AI agents?



LLMs + Tool chain



AI Agents

Image source: What are compound AI systems and AI agents?

Key challenges in Compound AI Systems

Many choices in Retrieval-Augmented Generation (RAG):

- Different retrieval and language models
- Techniques to improve retrieval (query expansion, reranking)
- Methods to refine LLM output (e.g., validation by another LLM)

Optimizing resource allocation:

- Balancing latency and cost among system components

Key challenges in Compound AI Systems

Compound systems require co-optimization:

- Example: In RAG, an LLM generates queries for a retriever and then produces an answer.
- Optimizing both the LLM and retriever together improves system performance.

Challenges vs. Single-Model Optimization:

- Traditional ML (e.g., PyTorch) allows easy end-to-end optimization.
- AI systems include **non-differentiable** components (e.g., search engines, code interpreters).

Key challenges in Compound AI Systems

Operation

Tracking & debugging become harder:

- Traditional ML (e.g., spam classifier) has clear success metrics.
- LLM agents may use **variable steps** (e.g., reflections, API calls), complicating evaluation.

Key areas for next-gen MLOps tools:

- **Monitoring:** Efficiently logging, analyzing, and debugging AI system traces.
- **DataOps:** Managing data pipelines for components like vector DBs, ensuring data quality.
- **Security:** Addressing risks unique to compound AI (e.g., LLM + content filter vulnerabilities, PII).

conclusion

- Small models in system settings >> Large language models.
- Build AI systems collaboratively (AI Scientists + AI engineers).
- Build workflows or Agentic Approaches for complex problems.
- Experiment with different libraries, review their code base.
- Always Evaluate!

Some Library recommendations:
Prompting: DSPY
Building Agents: LangGraph, SmolAgent, AutoGen

Recommended: Review selection of libraries at **llm-engineer-toolkit** by KalyanKS

Source

- Stanford Webinar - Large Language Models Get the Hype, but Compound Systems Are the Future of AI, [video]
- The Shift from Models to Compound AI Systems, Zaharia et al, [blog]
- What are compound AI systems and AI agents? By Microsoft, [blog]