Agent Fundamentals

Chen-Yu Lee (chenyulee@google.com)

Google

Expo Talk Panel

The Co-X Framework: Versatile AI Agents for Automating and Augmenting Professional Workflows

Chen-Yu Lee · Jinsung Yoon · Yale Song · Tomas Pfister

Upper Level Ballroom 20D

[Abstract]

Wed 3 Dec 4:30 p.m. PST — 5:30 p.m. PST (Bookmark)


Gemini Enterprise

Agent Fundamentals

Deep Research

Agent Designer

Gemini Enterprise

Bring the best of Google AI to every employee, for every workflow.

The Brains

Immediate access to Google's most advanced Gemini models

The Taskforce

A suite of specialized, Google-built & third-party agents from research to coding

The Workbench

Gemini chat platform with tools for all employees to build, orchestrate & use agents

The Context

All grounded in the reality of your own systems and data, wherever they live

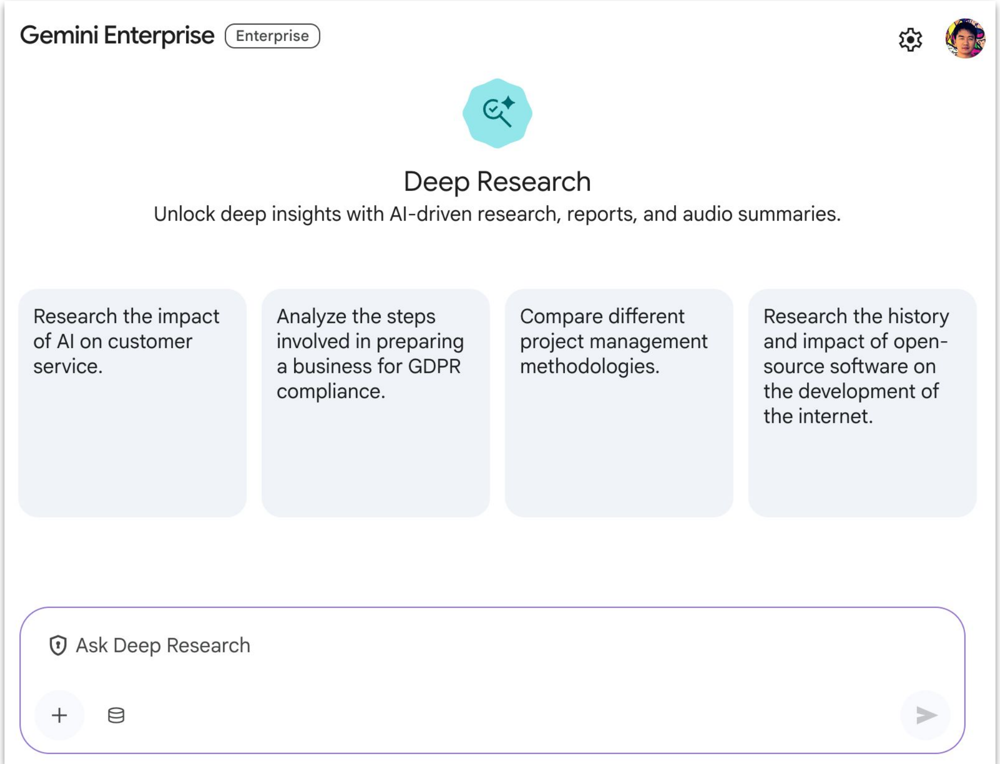
and many more...

Agent Memory

01

Deep Research

What is Deep Research?



An AI agent that automates the research lifecycle, including searching, reading, reasoning, and reporting, to facilitate the learning of new topics or skills.

Sample output

Step 1: query decomposition

Query: What are the trends in LLM research from NeurIPS 2025 papers? Given these trends, what should I research in LLMs as a PhD student in machine learning?

Research Plan: LLM Research Trends and PhD Opportunities

Analysis of LLM Research at NeurIPS 2025

I will begin by conducting a thorough analysis of the accepted papers at the NeurIPS 2025 conference to identify the most prominent research trends in Large Language Models. My research will focus on:

- **Categorizing LLM-related papers** into key themes such as reasoning, multimodality, efficiency and scaling, reinforcement learning and alignment, explainability, and ethical AI.
- **Identifying novel techniques and model architectures** presented in the papers, with a particular focus on methods that improve model performance, efficiency, and reliability.
- **Analyzing the focus of spotlight and oral presentations** to gauge which topics the community considers most significant.

Broader Trends in LLM Research

Next, I will broaden the scope of the research to understand the general direction of LLM research beyond the NeurIPS 2025 conference. This will involve:

- **Investigating the key challenges** that researchers are currently trying to address, such as reducing hallucinations, improving fact-checking, and mitigating biases.
- **Exploring the development of new capabilities** for LLMs, including the creation of autonomous agents and the use of synthetic data for training.
- **Examining the application of LLMs in other scientific disciplines** to understand how they are being used to accelerate research and discovery.

Promising Research Directions for a PhD Student

Finally, I will synthesize the findings from the previous sections to propose specific and promising research directions for a PhD student in machine learning. I will focus on identifying areas that are both impactful and feasible for academic research, including:

- **Novel architectures and training methods** that offer significant improvements in efficiency and performance, which is particularly relevant for academic labs with limited computational resources.
- **Advanced reasoning and planning techniques**, such as hierarchical planning, which are still in their early stages of development and offer many opportunities for fundamental contributions.
- **Explainability and interpretability of LLMs**, a critical area of research to build trust and understanding of these complex models.
- **The theoretical foundations of LLMs**, for students with a strong mathematical background, to develop a deeper understanding of why these models work and how to improve them.
- **Specialized applications of LLMs** in a specific domain of interest, which can lead to high-impact research and practical applications.

Analysis of LLM Research at NeurIPS 2025

Broader Trends in LLM Research

Promising Research Directions for a PhD Student

Sample output

Step 2: search question generation

Research details

1 Based on the papers from the NeurIPS 2023 and 2025 conferences, what were the dominant research trends and significant advancements in methods for improving the efficiency and scaling of Large Language Models, covering aspects like model compression, training...

Sample output

Research details

1 Based on the papers from the NeurIPS 2023 and 2025 conferences, what were the dominant research trends and significant advancements in methods for improving the efficiency and scaling of Large Language Models, covering aspects like model compression, training...

Step 3: answer searching

Step 4: reading + reasoning

Step 5: output drafting (denoising)

Sample output

Step 6: repeat step 2-5 for test-time scaling (iter 2)

Research details

1. Based on the papers from the NeurIPS 2023 and 2025 conferences, what were the dominant research trends and significant advancements in methods for improving the efficiency and scaling of Large Language Models, covering aspects like model compression, training...

2. Based on the NeurIPS 2025 award-winning paper "Artificial Hivemind: The Open-Ended Homogeneity of Language Models," provide a detailed analysis of the "homogenization" problem, explaining the methodology used to demonstrate this phenomenon, its proposed...

Draft from iter 1

Sample output

Continue test-time scaling
with desired # of iterations

Research details

1. Based on the papers from the NeurIPS 2023 and 2025 conferences, what were the dominant research trends and significant advancements in methods for improving the efficiency and scaling of Large Language Models, covering aspects like model compression, training...

2. Based on the NeurIPS 2025 award-winning paper "Artificial Hivemind: The Open-Ended Homogeneity of Language Models," provide a detailed analysis of the "homogenization" problem, explaining the methodology used to demonstrate this phenomenon, its proposed...

3. Based on the research trends from NeurIPS 2025, provide a detailed analysis of the award-winning paper that successfully scaled a self-supervised reinforcement learning (RL) agent to over 1,000 layers, as part of the broader trend of using LLMs as reasoning engines for...

⋮

Sample output

Executive Summary

The LLM research landscape is shifting from raw scaling to a more nuanced focus on **efficiency, agentic capabilities, advanced reasoning, and fundamental safety**. A major overarching theme is the pivot to **data-centric AI**, where data quality and curation are now seen as primary drivers of performance howaiworks.ai +3 .

Key emerging trends include:

- **Architectural Alternatives & Efficiency:** The rise of **State-Space Models (SSMs)** like **Mamba** and **Routing Mamba (RoM)** offers a highly efficient, linear-time alternative to Transformers, using a **Selective Scan Mechanism (S6)** for content-based reasoning howaiworks.ai +4 . In parallel, methods like **QLoRA** are democratizing research by enabling massive model fine-tuning on single GPUs through techniques like **4-bit NormalFloat (NF4)** and **Double Quantization** twosigma.com +4 .
- **Advanced Agents and Reasoning Evaluation:** Breakthroughs in agentic AI, such as the **"Physics Supernova"** agent, demonstrate elite human-level performance on complex scientific tasks by using a **"Reason-Act" loop** and specialized tools relational.ai +3 . This progress is measured by new, challenging benchmarks like the **PHYSICS dataset** arxiv.org +3 .
- **A New Generation of Safety Research and Benchmarks:** A critical concern is the **"Artificial Hivemind"**—the homogenization of LLM responses due to current alignment techniques medium.com +3 . To ensure reliability, the field is developing sophisticated new benchmarks like **Infinity-Chat** (for diversity), **DeceptionBench** (for detecting deception), **AbsenceBench** (for identifying missing information), and **VivaBench** (for interactive reasoning) vessl.ai +9 .
- **Data-Centric AI:** With high-quality text data becoming a bottleneck, the focus is on meticulous **data curation and optimization** relational.ai . The **NeurIPS LLM Efficiency Challenge** proved data strategy was key vessl.ai +2 , while new research provides **principled laws for optimal data mixing** youtube.com and scaling in data-constrained settings twosigma.com +2 .

For a PhD student, this opens up incredible opportunities. Instead of competing on raw model scale, you can make a name for yourself by:

1. **Designing more efficient non-Transformer or hybrid architectures like next-generation SSMs.**
2. **Tackling fundamental safety problems like the "Artificial Hivemind" by developing new alignment techniques that promote diversity.**
3. **Creating models with provable reasoning skills and building systems to mitigate novel failure modes identified by benchmarks like DeceptionBench and AbsenceBench.**
4. **Pioneering new methods in data curation, filtering, and composition, which is now a central pillar of AI research.**

The field is wide open for foundational contributions. My advice is to pick a direction that genuinely excites you and dive in. Good luck

Key emerging trends

- Architectural Alternatives & Efficiency
- Advanced Agents and Reasoning
- A New Generation of Safety Research and Benchmarks
- Data-Centric AI

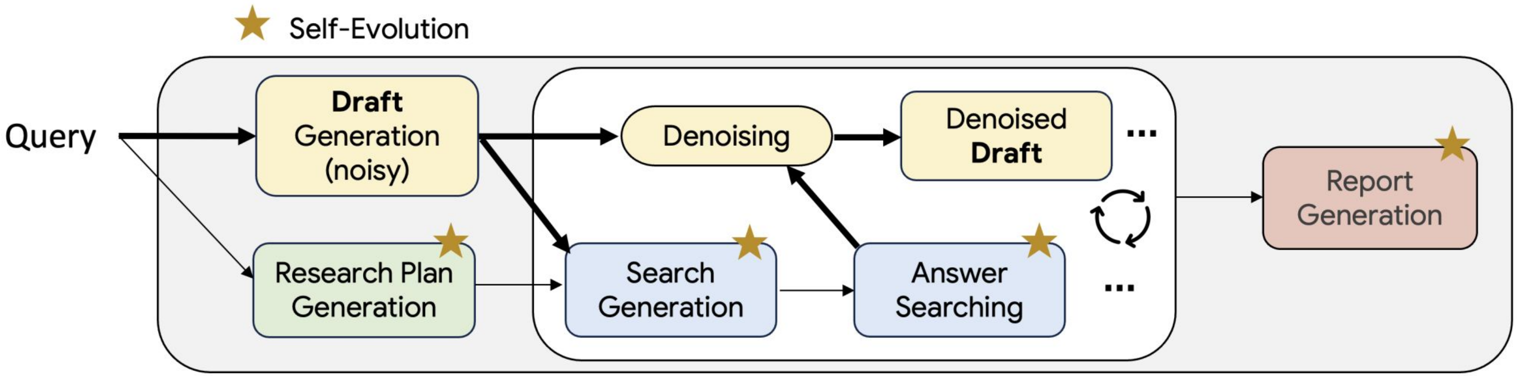
For a PhD student

- Designing more efficient non-Transformer or hybrid architectures like next-generation State-Space Models (SSMs)
- Tackling fundamental safety problems like the "Artificial Hivemind"
- Creating models with provable reasoning skills
- Pioneering new methods in data curation, filtering, and composition

Our method: Deep Researcher with Test-Time Diffusion

Our method is inspired by the natural human writing process, which includes planning, drafting, and multiple revisions to the draft

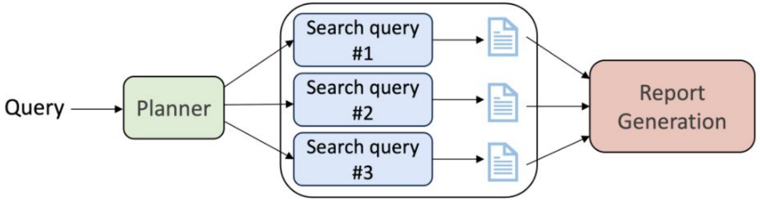
Our method: Deep Researcher with Test-Time Diffusion



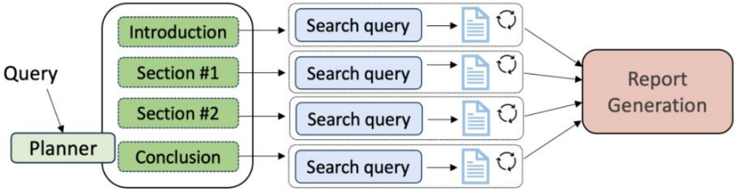
This **draft-centric design** makes the report writing process more timely and coherent while reducing information loss during the iterative search process

Comparisons with other Deep Research methods

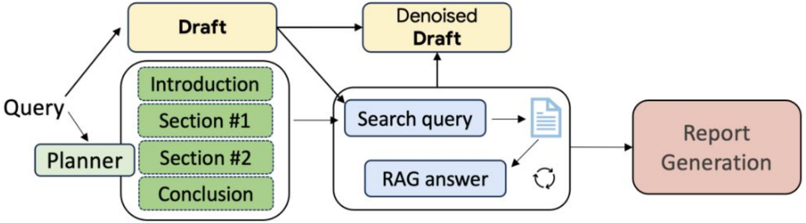
(a) Huggingface Open DR



(b) GPT Researcher



(c) Open Deep Research



(d) Test-Time Diffusion DR (ours)

Results

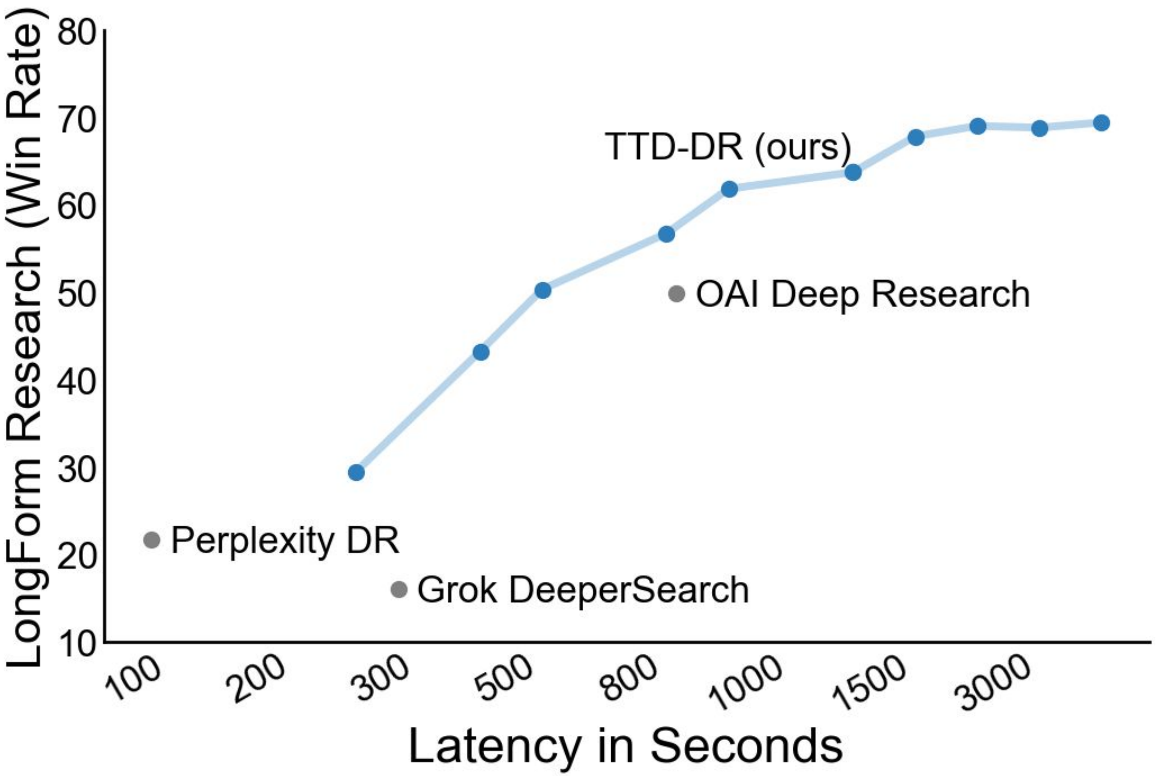


figure plotted in July 2025

Paper

Deep Researcher with Test-Time Diffusion

Rujun Han[*1], Yanfei Chen[*1], Zoey CuiZhu[2], Lesly Miculicich[1], Guan Sun[2], Yuanjun Bi[2], Weiming Wen[2], Hui Wan[2], Chunfeng Wen[2], Solène Maître[2], George Lee[1], Vishy Tirumalashetty[2], Emily Xue[2], Zizhao Zhang[2], Salem Haykal[2], Burak Gokturk[1], Tomas Pfister[1] and Chen-Yu Lee[1]
[1]Google Cloud AI Research, [2]Google Cloud

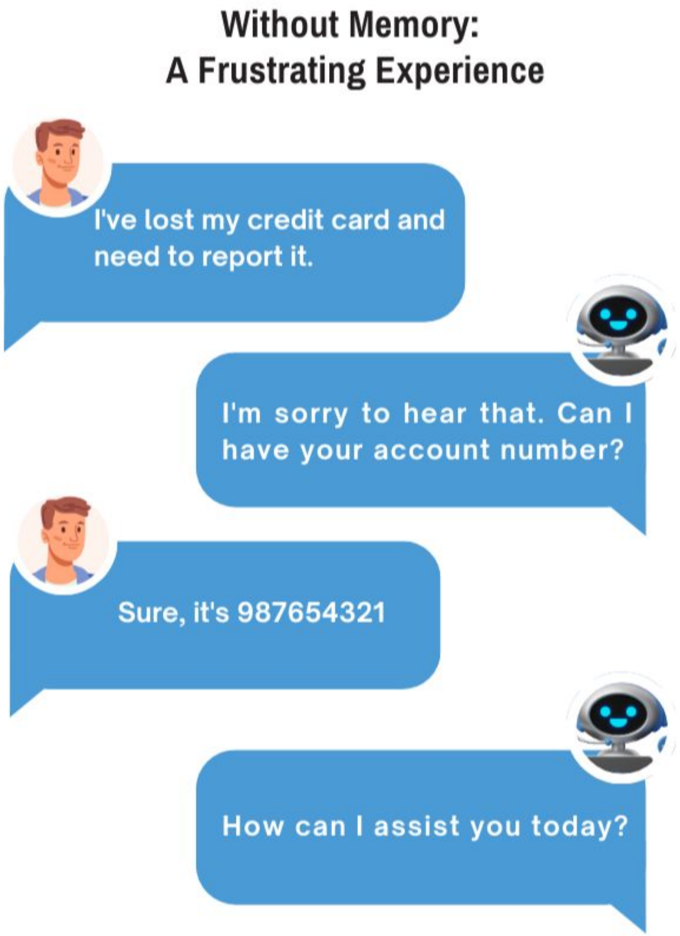
Deep research agents, powered by Large Language Models (LLMs), are rapidly advancing; yet, their performance often plateaus when generating complex, long-form research reports using generic test-time scaling algorithms. Drawing inspiration from the iterative nature of human research, which involves cycles of searching, reasoning, and revision, we propose the Test-Time Diffusion Deep Researcher (TTD-DR). This novel framework conceptualizes research report generation as a diffusion process. TTD-DR initiates this process with a preliminary draft, an updatable skeleton that serves as an evolving foundation to guide the research direction. The draft is then iteratively refined through a "denoising" process, which is dynamically informed by a retrieval mechanism that incorporates external information at each step. The core process is further enhanced by a self-evolutionary algorithm applied to each component of the agentic workflow, ensuring the generation of high-quality context for the diffusion process. This draft-centric design makes the report writing process more timely and coherent while reducing information loss during the iterative search process. We demonstrate that our TTD-DR achieves state-of-the-art results on a wide array of benchmarks that require intensive search and multi-hop reasoning, significantly outperforming existing deep research agents.

02

Agent Memory

Ever talk to a forgetful agent?

- Agents ask the same questions over and over
- Agents don't remember who you are
- Every chat session starts from scratch



Sometimes it's more than frustrating

Session 1 (3 days ago)

I have a persistent cough and a slight fever. — Patient

Healthcare Agent: Okay. Please tell me about the severity of each symptom.

⋮

Session 2 (Today)

I have a headache and the fever is gone. — Patient

with Memory

Healthcare Agent: Thanks. So, to update: **the fever is gone, a headache started**, and **the cough persists.** Given that progression, let's consider ...

without Memory

Okay. Given the headache, let's consider ...

Why not feed all past interactions into LLM context?

- Higher cost
- Degraded performance

3P Model	Accuracy (%)
No Memory	10.8
Full Context	55.4
RMB (ours w/o re-ranker)	**74.6**

RMB: Reflective Memory Bank

Proposal 1: MemoryBank

In Prospect and Retrospect: Reflective Memory Management for Long-term Personalized Dialogue Agents

Zhen Tan[1 *], Jun Yan[2], I-Hung Hsu[2], Rujun Han[2], Zifeng Wang[2], Long T. Le[2], Yiwen Song[2], Yanfei Chen[2], Hamid Palangi[2], George Lee[2], Anand Iyer[3], Tianlong Chen[4], Huan Liu[1], Chen-Yu Lee[2] and Tomas Pfister[2]
[1]Arizona State University, [2]Google Cloud AI Research, [3]Google Cloud AI, [4]UNC Chapel Hill

Large Language Models (LLMs) have made significant progress in open-ended dialogue, yet their inability to retain and retrieve relevant information from long-term interactions limits their effectiveness in applications requiring sustained personalization. External memory mechanisms have been proposed to address this limitation, enabling LLMs to maintain conversational continuity. However, existing approaches struggle with two key challenges. First, rigid memory granularity fails to capture the natural semantic structure of conversations, leading to fragmented and incomplete representations. Second, fixed retrieval mechanisms cannot adapt to diverse dialogue contexts and user interaction patterns. In this work, we propose Reflective Memory Management (RMM), a novel mechanism for long-term dialogue agents, integrating forward- and backward-looking reflections: (1) Prospective Reflection, which dynamically summarizes interactions across granularities—utterances, turns, and sessions—into a personalized memory bank for effective future retrieval, and (2) Retrospective Reflection, which iteratively refines the retrieval in an online reinforcement learning (RL) manner based on LLMs' cited evidence. Experiments show that RMM demonstrates consistent improvement across various metrics and benchmarks. For example, RMM shows more than 10% accuracy improvement over the baseline without memory management on the `LongMemEval` dataset.

ACL 2025

Topic-based Memory Management

- Prospective Reflection: Auto-topic division, key info summarization, and organized storage in memory after each chat.

- Retrospective Reflection: Retrieval of potentially relevant topic summaries from memory upon new user interaction.

Proposal 2: ReasoningBank

Google

ReasoningBank: Scaling Agent Self-Evolving with Reasoning Memory

Siru Ouyang[1*], Jun Yan[2], I-Hung Hsu[2], Yanfei Chen[2], Ke Jiang[2], Zifeng Wang[2], Rujun Han[2], Long T. Le[2], Samira Daruki[2], Xiangru Tang[3], Vishy Tirumalashetty[2], George Lee[2], Mahsan Rofouei[4], Hangfei Lin[4], Jiawei Han[1], Chen-Yu Lee[2] and Tomas Pfister[2]

[1]University of Illinois Urbana-Champaign, [2]Google Cloud AI Research, [3]Yale University, [4]Google Cloud AI

With the growing adoption of large language model agents in persistent real-world roles, they naturally encounter continuous streams of tasks. A key limitation, however, is their failure to learn from the accumulated interaction history, forcing them to discard valuable insights and repeat past errors. We propose ReasoningBank, a novel memory framework that distills generalizable reasoning strategies from an agent's self-judged successful and failed experiences. At test time, an agent retrieves relevant memories from ReasoningBank to inform its interaction and then integrates new learnings back, enabling it to become more capable over time. Building on this powerful experience learner, we further introduce memory-aware test-time scaling (MaTTS), which accelerates and diversifies this learning process by scaling up the agent's interaction experience. By allocating more compute to each task, the agent generates abundant, diverse experiences that provide rich contrastive signals for synthesizing higher-quality memory. The better memory in turn guides more effective scaling, establishing a powerful synergy between memory and test-time scaling. Across web browsing and software engineering benchmarks, ReasoningBank consistently outperforms existing memory mechanisms that store raw trajectories or only successful task routines, improving both effectiveness and efficiency; MaTTS further amplifies these gains. These findings establish *memory-driven experience scaling* as a new scaling dimension, enabling agents to self-evolve with emergent behaviors naturally arise.

- Beyond memorizing user-agent interactions, but also agent-environment interactions
- Distill task insights from both successful and failed experiences

Available in Gemini Enterprise & Google Cloud Vertex AI

Google Cloud

Contact sales | Get started for free

Blog | Solutions & technology | Ecosystem | Developers & Practitioners | Transform with Google Cloud

AI & Machine Learning

Announcing Vertex AI Agent Engine Memory Bank available for everyone in preview

July 8, 2025

Kimberly Milam
Software Engineer, Vertex AI

George Lee
Product Manager, Cloud AI Research

Developers are racing to productize agents, but a common limitation is the absence of memory. Without memory, agents treat each interaction as the first, asking repetitive questions and failing to recall user preferences. This lack of contextual awareness makes it difficult for an agent to personalize their assistance–and leaves developers frustrated.

How we normally mitigate memory problems: So far, a common approach to this problem has been to leverage the LLM's context window. However, directly inserting entire session dialogues into an LLM's context window is both expensive and computationally inefficient, leading to higher inference costs and slower response times. Also, as the amount of information fed into an LLM grows, especially with irrelevant or misleading details, the quality of the model's output significantly declines, leading to issues like "lost in the middle" and "context rot".

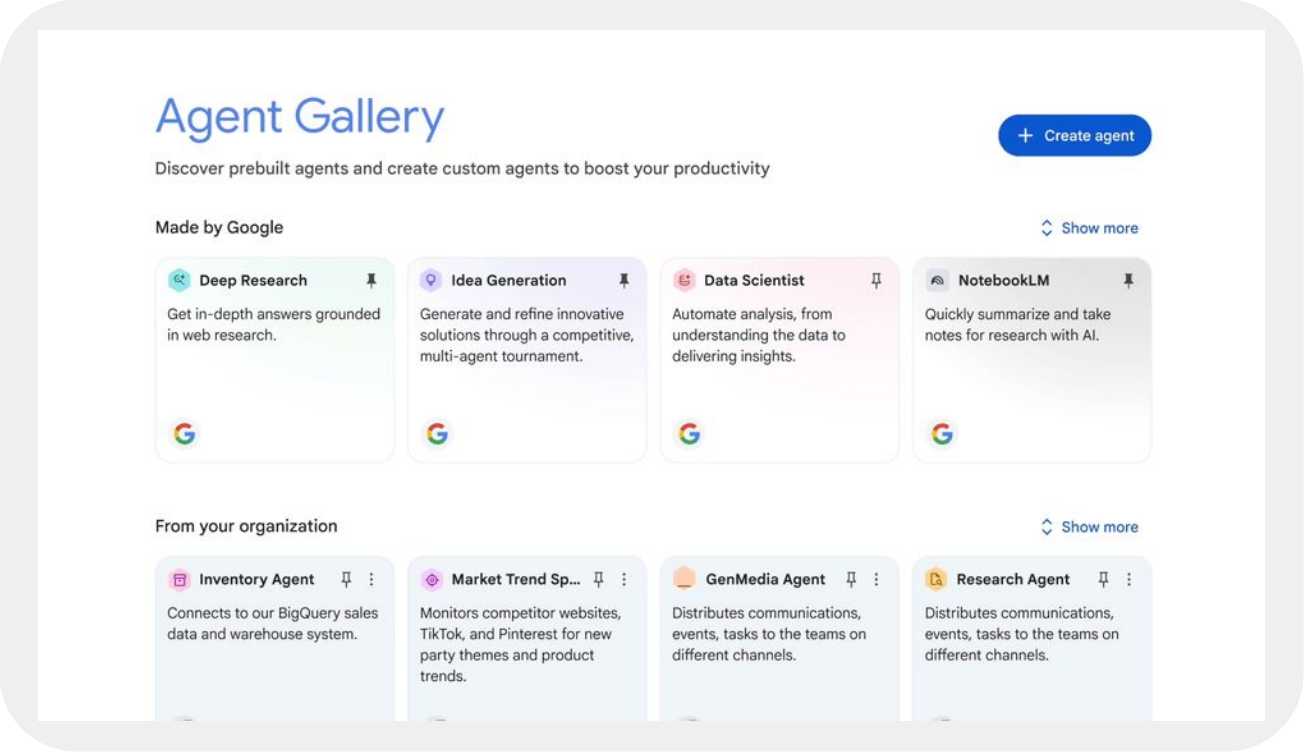
03

Agent Designer

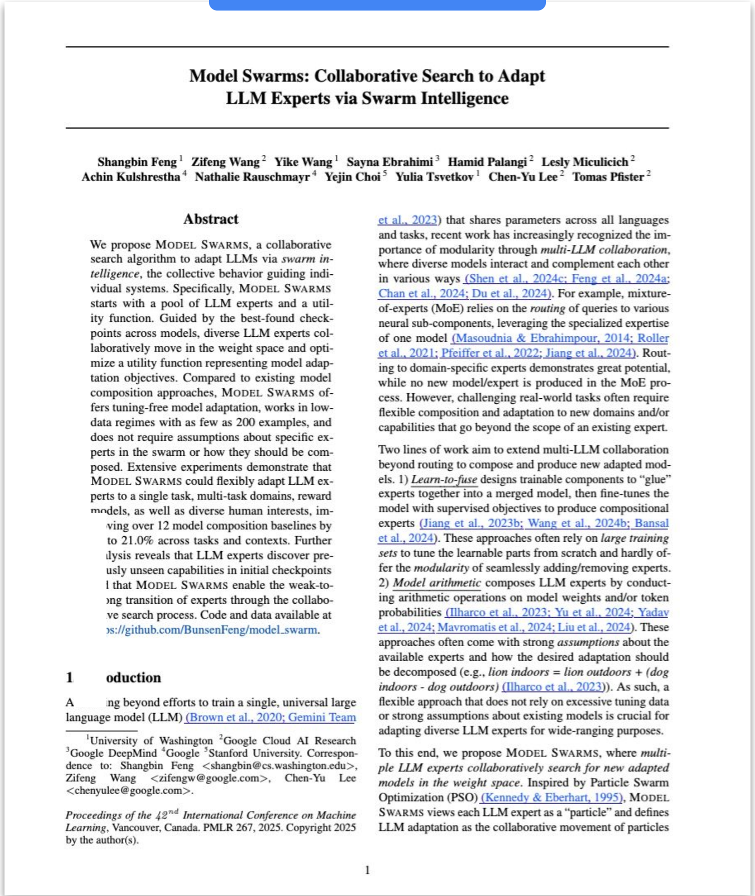
Agent Designer

No-code Agent Builder

- Build ADK Agents via chat interface
- Gemini-assistant to build agents
- 10+ connectors supported, for example
  - **1p**: Gmail, G-Drive, G-Calendar
  - **3p**: Confluence, Jira
- Multi-agent generation supported



Model Swarms



Model Swarms: Collaborative Search to Adapt LLM Experts via Swarm Intelligence

Shangbin Feng [1] Zifeng Wang [2] Yike Wang [1] Sayna Ebrahimi [3] Hamid Palangi [2] Lesly Miculicich [2] Achin Kulshrestha [4] Nathalie Rauschmayr [4] Yejin Choi [5] Yulia Tsvetkov [1] Chen-Yu Lee [2] Tomas Pfister [2]

Abstract

We propose MODEL SWARMS, a collaborative search algorithm to adapt LLMs via *swarm intelligence*, the collective behavior guiding individual systems. Specifically, MODEL SWARMS starts with a pool of LLM experts and a utility function. Guided by the best-found checkpoints across models, diverse LLM experts collaboratively move in the weight space and optimize a utility function representing model adaptation objectives. Compared to existing model composition approaches, MODEL SWARMS offers tuning-free model adaptation, works in low-data regimes with as few as 200 examples, and does not require assumptions about specific experts in the swarm or how they should be composed. Extensive experiments demonstrate that MODEL SWARMS could flexibly adapt LLM experts to a single task, multi-task domains, reward models, as well as diverse human interests, im- ving over 12 model composition baselines by to 21.0% across tasks and contexts. Further lysis reveals that LLM experts discover pre- usly unseen capabilities in initial checkpoints l that MODEL SWARMS enable the weak-to- ong transition of experts through the collabo- ve search process. Code and data available at s://github.com/BunsenFeng/model_swarm.

1 oduction

A ng beyond efforts to train a single, universal large language model (LLM) (Brown et al., 2020; Gemini Team et al., 2023) that shares parameters across all languages and tasks, recent work has increasingly recognized the importance of modularity through *multi-LLM collaboration*, where diverse models interact and complement each other in various ways (Shen et al., 2024c; Feng et al., 2024a; Chan et al., 2024; Du et al., 2024). For example, mixture-of-experts (MoE) relies on the *routing* of queries to various neural sub-components, leveraging the specialized expertise of one model (Masoudnia & Ebrahimpour, 2014; Roller et al., 2021; Pfeiffer et al., 2022; Jiang et al., 2024). Routing to domain-specific experts demonstrates great potential, while no new model/expert is produced in the MoE process. However, challenging real-world tasks often require flexible composition and adaptation to new domains and/or capabilities that go beyond the scope of an existing expert.

Two lines of work aim to extend multi-LLM collaboration beyond routing to compose and produce new adapted models. 1) *Learn-to-fuse* designs trainable components to "glue" experts together into a merged model, then fine-tunes the model with supervised objectives to produce compositional experts (Jiang et al., 2023b; Wang et al., 2024b; Bansal et al., 2024). These approaches often rely on *large training sets* to tune the learnable parts from scratch and hardly offer the *modularity* of seamlessly adding/removing experts. 2) *Model arithmetic* composes LLM experts by conducting arithmetic operations on model weights and/or token probabilities (Ilharco et al., 2023; Yu et al., 2024; Yadav et al., 2024; Mavromatis et al., 2024; Liu et al., 2024). These approaches often come with strong *assumptions* about the available experts and how the desired adaptation should be decomposed (e.g., *lion indoors = lion outdoors + (dog indoors - dog outdoors)* (Ilharco et al., 2023)). As such, a flexible approach that does not rely on excessive tuning data or strong assumptions about existing models is crucial for adapting diverse LLM experts for wide-ranging purposes.

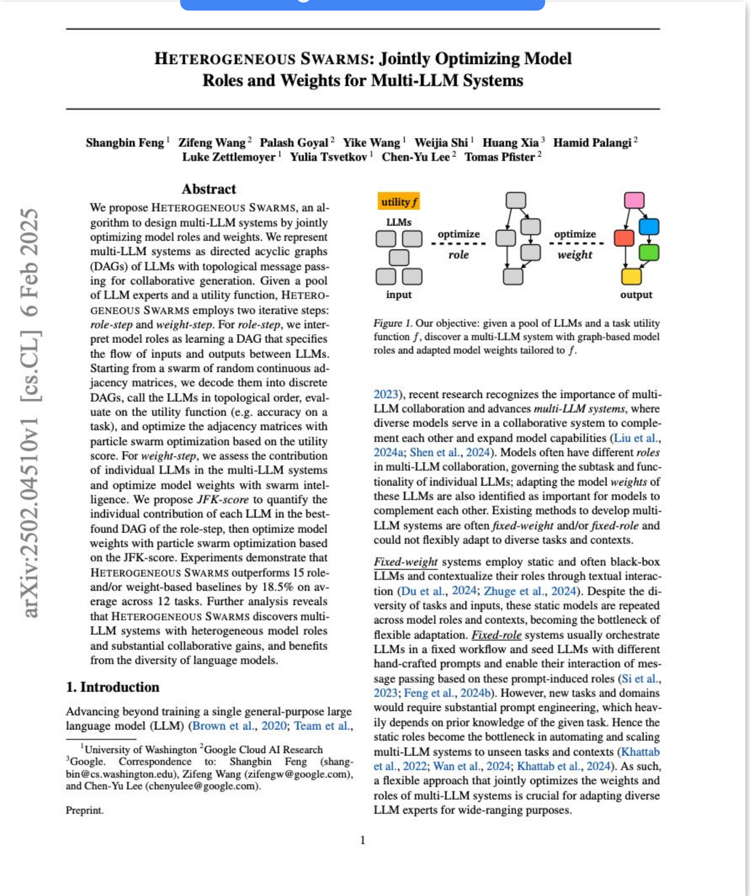
To this end, we propose MODEL SWARMS, where *multiple LLM experts collaboratively search for new adapted models in the weight space*. Inspired by Particle Swarm Optimization (PSO) (Kennedy & Eberhart, 1995), MODEL SWARMS views each LLM expert as a "particle" and defines LLM adaptation as the collaborative movement of particles

[1]University of Washington [2]Google Cloud AI Research [3]Google DeepMind [4]Google [5]Stanford University. Correspondence to: Shangbin Feng <shangbin@cs.washington.edu>, Zifeng Wang <zifengw@google.com>, Chen-Yu Lee <chenyulee@google.com>.

Proceedings of the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

ICML 2025

Heterogeneous Swarms



HETEROGENEOUS SWARMS: Jointly Optimizing Model Roles and Weights for Multi-LLM Systems

Shangbin Feng [1] Zifeng Wang [2] Palash Goyal [2] Yike Wang [1] Weijia Shi [1] Huang Xia [3] Hamid Palangi [2] Luke Zettlemoyer [1] Yulia Tsvetkov [1] Chen-Yu Lee [2] Tomas Pfister [2]

Abstract

We propose HETEROGENEOUS SWARMS, an algorithm to design multi-LLM systems by jointly optimizing model roles and weights. We represent multi-LLM systems as directed acyclic graphs (DAGs) of LLMs with topological message passing for collaborative generation. Given a pool of LLM experts and a utility function, HETEROGENEOUS SWARMS employs two iterative steps: *role-step* and *weight-step*. For *role-step*, we interpret model roles as learning a DAG that specifies the flow of inputs and outputs between LLMs. Starting from a swarm of random continuous adjacency matrices, we decode them into discrete DAGs, call the LLMs in topological order, evaluate on the utility function (e.g. accuracy on a task), and optimize the adjacency matrices with particle swarm optimization based on the utility score. For *weight-step*, we assess the contribution of individual LLMs in the multi-LLM systems and optimize model weights with swarm intelligence. We propose *JFK-score* to quantify the individual contribution of each LLM in the best-found DAG of the role-step, then optimize model weights with particle swarm optimization based on the JFK-score. Experiments demonstrate that HETEROGENEOUS SWARMS outperforms 15 role- and/or weight-based baselines by 18.5% on average across 12 tasks. Further analysis reveals that HETEROGENEOUS SWARMS discovers multi-LLM systems with heterogeneous model roles and substantial collaborative gains, and benefits from the diversity of language models.

Figure 1. Our objective: given a pool of LLMs and a task utility function f, discover a multi-LLM system with graph-based model roles and adapted model weights tailored to f.

1. Introduction

Advancing beyond training a single general-purpose large language model (LLM) (Brown et al., 2020; Team et al., 2023), recent research recognizes the importance of multi-LLM collaboration and advances *multi-LLM systems*, where diverse models serve in a collaborative system to complement each other and expand model capabilities (Liu et al., 2024a; Shen et al., 2024). Models often have different *roles* in multi-LLM collaboration, governing the subtask and functionality of individual LLMs; adapting the model *weights* of these LLMs are also identified as important for models to complement each other. Existing methods to develop multi-LLM systems are often *fixed-weight* and/or *fixed-role* and could not flexibly adapt to diverse tasks and contexts.

Fixed-weight systems employ static and often black-box LLMs and contextualize their roles through textual interaction (Du et al., 2024; Zhuge et al., 2024). Despite the diversity of tasks and inputs, these static models are repeated across model roles and contexts, becoming the bottleneck of flexible adaptation. *Fixed-role* systems usually orchestrate LLMs in a fixed workflow and seed LLMs with different hand-crafted prompts and enable their interaction of message passing based on these prompt-induced roles (Si et al., 2023; Feng et al., 2024b). However, new tasks and domains would require substantial prompt engineering, which heavily depends on prior knowledge of the given task. Hence the static roles become the bottleneck in automating and scaling multi-LLM systems to unseen tasks and contexts (Khattab et al., 2022; Wan et al., 2024; Khattab et al., 2024). As such, a flexible approach that jointly optimizes the weights and roles of multi-LLM systems is crucial for adapting diverse LLM experts for wide-ranging purposes.

[1]University of Washington [2]Google Cloud AI Research [3]Google. Correspondence to: Shangbin Feng (shangbin@cs.washington.edu), Zifeng Wang (zifengw@google.com), and Chen-Yu Lee (chenyulee@google.com).

Preprint.

NeurIPS 2025

Agent Fundamentals

Deep Research

Agent Designer

Gemini Enterprise

Bring the best of Google AI to every employee, for every workflow.

The Brains

Immediate access to Google's most advanced Gemini models

The Taskforce

A suite of specialized, Google-built & third-party agents from research to coding

The Workbench

Gemini chat platform with tools for all employees to build, orchestrate & use agents

The Context

All grounded in the reality of your own systems and data, wherever they live

and many more...

Agent Memory