Shakil Mahmud • Research Monograph

SHAKIL MAHMUD EDITORIAL & RESEARCH MONOGRAPHS

# লার্জ ল্যাঙ্গুয়েজ মডেল ও জেনারেটিভ এআই

*টোকেনাইজেশন ও স্কেলিং সূত্র থেকে মেগাট্রন ডিস্ট্রিবিউটেড ক্লাস্টার, ফ্ল্যাশ-অ্যাটেনশন, লোরা, আরএলএইচএফ, ডিপিও, ডিফিউশন এসডিই এবং মাল্টিমোডাল মডেলের ২০ খণ্ডের আধুনিক প্রযুক্তির মহাকাব্য।*

SHAKIL MAHMUD

*গবেষক ও লেখক*

SHAKIL MAHMUD EDITORIAL & RESEARCH SERIES

প্রথম সংকলিত সংস্করণ • ঢাকা ও সিঙ্গাপুর • ২০২৬

# লার্জ ল্যাঙ্গুয়েজ মডেল ও জেনারেটিভ এআই

*টোকেনাইজেশন ও স্কেলিং সূত্র থেকে মেগাট্রন ডিস্ট্রিবিউটেড ক্লাস্টার, ফ্ল্যাশ-অ্যাটেনশন, লোরা, আরএলএইচএফ, ডিপিও, ডিফিউশন এসডিই এবং মাল্টিমোডাল মডেলের ২০ খণ্ডের আধুনিক প্রযুক্তির মহাকাব্য।*

**LIBRARY & CATALOGING-IN-PUBLICATION DATA**

**Title:** লার্জ ল্যাঙ্গুয়েজ মডেল ও জেনারেটিভ এআই
**Author:** Shakil Mahmud
**Series:** Shakil Mahmud Editorial & Research Series (Monograph Collection)
**Subjects:** Science -- History -- Philosophy -- Mathematics -- Civilization
**Published:** Dhaka & Singapore, 2026
**Format:** Digital Monograph Offprint (PDF) • Complete 20-Part Collected Edition
**Permanent URI:** https://shakilmahmud.com/series/generative-ai/

**প্রস্তাবিত সাইটেশন:**
Mahmud, Shakil. (2026). *লার্জ ল্যাঙ্গুয়েজ মডেল ও জেনারেটিভ এআই: The Complete 20-Part Collected Monograph*. Shakil Mahmud Research Series. https://shakilmahmud.com/series/generative-ai/

**Open Scholarly Access**: This collected monograph is published under Open Scholarly Access for personal study, academic citation, university pedagogy, and non-commercial research.

© 2026 Shakil Mahmud. All rights reserved. Typeset and published at Dhaka and Singapore.

Shakil Mahmud • Research Monograph

তাত্ত্বিক ভূমিকা

# ভূমিকা: জ্ঞানের একীভূত রূপরেখা

❖

জেনারেটিভ এআই এবং লার্জ ল্যাঙ্গুয়েজ মডেল হলো মানুষের প্রতীকী চিন্তাশক্তির সফল গণনামুখী রূপায়ণ। সাবওয়ার্ড বাইট-পেয়ার টোকেনাইজেশন, নিউরাল স্কেলিং ল, ফ্ল্যাশ-অ্যাটেনশন, স্পার্স মিক্সচার-অব-এক্সপার্টস (MoE) থেকে লো-র‍্যাঙ্ক অ্যাডাপ্টেশন (LoRA), মানব প্রতিক্রিয়া ভিত্তিক রিইনফোর্সমেন্ট লার্নিং (RLHF), ডাইরেক্ট প্রেফারেন্স অপ্টিমাইজেশন (DPO) এবং অবিচ্ছিন্ন ডিফিউশন মডেল—বিশ খণ্ডে সংকলিত এই গবেষণা মনোগ্রাফ আধুনিক ফাউন্ডেশন মডেলের পূর্ণাঙ্গ কারিগরি রূপরেখা।

এই গবেষণামূলক কাজের উদ্দেশ্য কেবল তথ্য উপস্থাপন নয়, বরং চিন্তার গভীর বিকাশকে পাঠকের সামনে উন্মোচন করা। প্রতিটি পরিচ্ছদ এক একটি যুগান্তকারী বৌদ্ধিক সংকটের সমাধান হিসেবে কীভাবে আবির্ভূত হয়েছে, তা এখানে পদ্ধতিগতভাবে লিপিবদ্ধ করা হলো।

*— Shakil Mahmud*
*Dhaka & Singapore, 2026*

অধ্যায় সূচি

# সূচিপত্র

## প্রথম পর্যায়: টোকেনাইজেশন ও কম্পিউট স্কেলিং ল

খণ্ড ১ টোকেনাইজেশন বটলনেক — বাইট-পেয়ার এনকোডিং, ওয়ার্ডপিস ও সাবওয়ার্ড শব্দভাণ্ডার 353 শব্দ
*টোকেনাইজেশন বটলনেক — বাইট-পেয়ার এনকোডিং, ওয়ার্ডপিস ও সাবওয়ার্ড শব্দভাণ্ডার*

খণ্ড ২ নিউরাল স্কেলিং ল — কাপলান বনাম চিনচিলা কম্পিউট অপ্টিমালিটি 326 শব্দ
*নিউরাল স্কেলিং ল — কাপলান বনাম চিনচিলা কম্পিউট অপ্টিমালিটি*

খণ্ড ৩ ডেনস ট্রান্সফরমার আর্কিটেকচার — জিপিটি-৩, পালম, এলএলএএমএ ও আধুনিক রেসিপি 283 শব্দ
*ডেনস ট্রান্সফরমার আর্কিটেকচার — জিপিটি-৩, পালম, এলএলএএমএ ও আধুনিক রেসিপি*

খণ্ড ৪ কেভি ক্যাশ ও ইনফারেন্স লেটেন্সি — পেজড-অ্যাটেনশন, ভিএলএলএম ও সার্ভিং মেমরি 291 শব্দ
*কেভি ক্যাশ ও ইনফারেন্স লেটেন্সি — পেজড-অ্যাটেনশন, ভিএলএলএম ও সার্ভিং মেমরি*

## দ্বিতীয় পর্যায়: ফ্ল্যাশ-অ্যাটেনশন ও সিস্টেম আর্কিটেকচার

খণ্ড ৫ মাল্টি-কোয়েরি ও গ্রুপড-কোয়েরি অ্যাটেনশন — মেমরি ব্যান্ডউইথ সীমানা অতিক্রম 441 শব্দ
*মাল্টি-কোয়েরি ও গ্রুপড-কোয়েরি অ্যাটেনশন — মেমরি ব্যান্ডউইথ সীমানা অতিক্রম*

খণ্ড ৬ ফ্ল্যাশ-অ্যাটেনশন ও আই/ও সচেতনতা — জিপিইউ এস-র‍্যাম টাইলিং ও অনলাইন সফটম্যাক্স 398 শব্দ
*ফ্ল্যাশ-অ্যাটেনশন ও আই/ও সচেতনতা — জিপিইউ এস-র‍্যাম টাইলিং ও অনলাইন সফটম্যাক্স*

খণ্ড ৭ মডেল কোয়ান্টাইজেশন তত্ত্ব — FP16, INT8, INT4, GPTQ, AWQ ও আউটলায়ার ডাইমেনশন 325 শব্দ
*মডেল কোয়ান্টাইজেশন তত্ত্ব — FP16, INT8, INT4, GPTQ, AWQ ও আউটলায়ার ডাইমেনশন*

খণ্ড ৮ প্যারামিটার-দক্ষ ফাইন-টিউনিং — লোরা, কিউ-লোরা ও ইন্ট্রিনসিক র‍্যাংকের জ্যামিতি 365 শব্দ
*প্যারামিটার-দক্ষ ফাইন-টিউনিং — লোরা, কিউ-লোরা ও ইন্ট্রিনসিক র‍্যাংকের জ্যামিতি*

## তৃতীয় পর্যায়: স্পার্স এমওই ও ফাইন-টিউনিং অ্যাডাপ্টেশন

খণ্ড ৯ স্পার্স মিক্সচার-অব-এক্সপার্টস (MoE) — রাউটার গেট, লোড ব্যালান্সিং ও সুইচ ট্রান্সফরমার 364 শব্দ
*স্পার্স মিক্সচার-অব-এক্সপার্টস (MoE) — রাউটার গেট, লোড ব্যালান্সিং ও সুইচ ট্রান্সফরমার*

খণ্ড ১০ প্রি-ট্রেনিং গতিবিদ্যা ও ডেটা ইঞ্জিনিয়ারিং — মিনহাশ ডিডুপ্লিকেশন ও সিন্থেটিক ডেটাবেস 310 শব্দ
*প্রি-ট্রেনিং গতিবিদ্যা ও ডেটা ইঞ্জিনিয়ারিং — মিনহাশ ডিডুপ্লিকেশন ও সিন্থেটিক ডেটাবেস*

খণ্ড ১১ সুপারভাইজড ফাইন-টিউনিং ও ইন্সট্রাকশন টিউনিং — ক্রস-এনট্রপি গতিবিদ্যা 255 শব্দ
*সুপারভাইজড ফাইন-টিউনিং ও ইন্সট্রাকশন টিউনিং — ক্রস-এনট্রপি গতিবিদ্যা*

খণ্ড ১২ রিইনফোর্সমেন্ট লার্নিং ফ্রম হিউম্যান ফিডব্যাক (RLHF) — ব্র্যাডলি-টেরি রিওয়ার্ড মডেল ও পিপিও 314 শব্দ
*রিইনফোর্সমেন্ট লার্নিং ফ্রম হিউম্যান ফিডব্যাক (RLHF) — ব্র্যাডলি-টেরি রিওয়ার্ড মডেল ও পিপিও*

খণ্ড ১৩ ডাইরেক্ট প্রেফারেন্স অপটিমাইজেশন (DPO) — রিওয়ার্ড মডেলবিহীন ইমপ্লিসিট লগ-অডস 294 শব্দ
*ডাইরেক্ট প্রেফারেন্স অপটিমাইজেশন (DPO) — রিওয়ার্ড মডেলবিহীন ইমপ্লিসিট লগ-অডস*

খণ্ড ১৪ প্রম্পট ইঞ্জিনিয়ারিং ও ইন-কনটেক্সট লার্নিং — ইন্ডাকশন হেড ও চেইন-অব-থট যুক্তি 259 শব্দ
*প্রম্পট ইঞ্জিনিয়ারিং ও ইন-কনটেক্সট লার্নিং — ইন্ডাকশন হেড ও চেইন-অব-থট যুক্তি*

## চতুর্থ পর্যায়: অ্যালাইনমেন্ট, ডিফিউশন ও ফ্রন্টিয়ার সিস্টেম

খণ্ড ১৫ রিট্রিভাল-অগমেন্টেড জেনারেশন (RAG) — ডেন্স এম্বেডিং, HNSW ইনডেক্সিং ও হাইব্রিড রির‍্যাংকিং 271 শব্দ
*রিট্রিভাল-অগমেন্টেড জেনারেশন (RAG) — ডেন্স এম্বেডিং, HNSW ইনডেক্সিং ও হাইব্রিড রির‍্যাংকিং*

Shakil Mahmud • Research Monograph

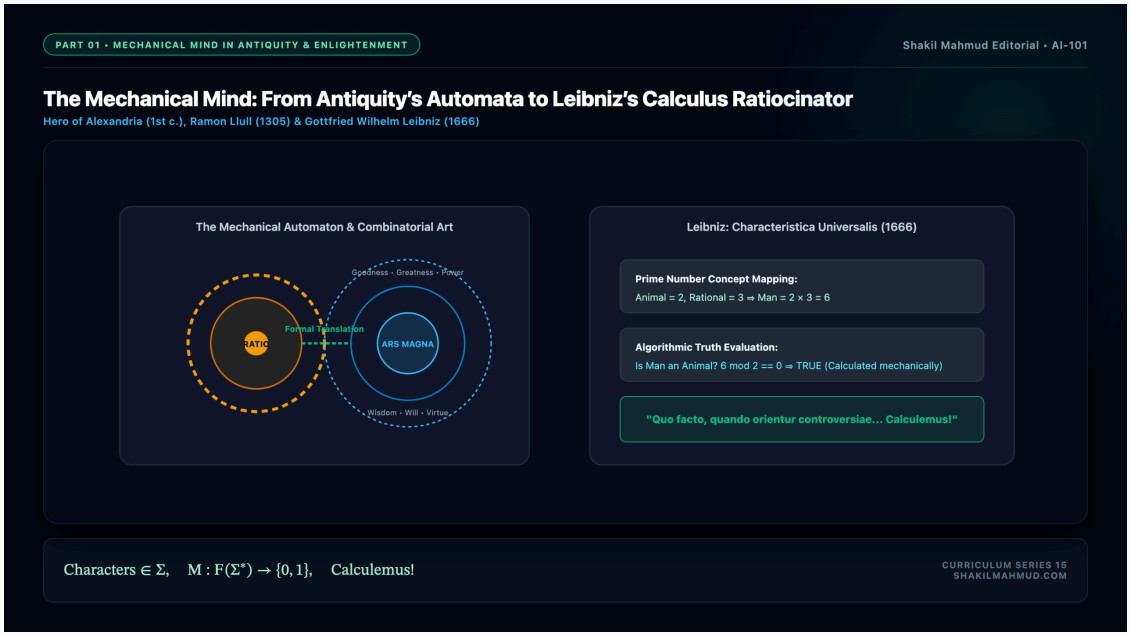
| | | |
|---|---|---|
| খণ্ড ১৬ | **অটোরিগ্রেসিভ ডিকোডিং কৌশল — টেম্পারেচার, টপ-পি নিউক্লিয়াস ও স্পেকুলেটিভ ডিকোডিং**<br>*অটোরিগ্রেসিভ ডিকোডিং কৌশল — টেম্পারেচার, টপ-পি নিউক্লিয়াস ও স্পেকুলেটিভ ডিকোডিং* | 311 শব্দ |
| খণ্ড ১৭ | **স্কোর-ভিত্তিক কন্টিনিউয়াস ডিফিউশন — স্টোকাস্টিক ডিফারেনশিয়াল সমীকরণ ও DDPM**<br>*স্কোর-ভিত্তিক কন্টিনিউয়াস ডিফিউশন — স্টোকাস্টিক ডিফারেনশিয়াল সমীকরণ ও DDPM* | 280 শব্দ |
| খণ্ড ১৮ | **ল্যাটেন্ট ডিফিউশন ও ফ্লো ম্যাচিং — স্টেবল ডিফিউশন ও রেকটিফায়েড ফ্লোস**<br>*ল্যাটেন্ট ডিফিউশন ও ফ্লো ম্যাচিং — স্টেবল ডিফিউশন ও রেকটিফায়েড ফ্লোস* | 265 শব্দ |
| খণ্ড ১৯ | **লং-কনটেক্সট সিস্টেম — রিং-অ্যাটেনশন, RoPE ফ্রিকোয়েন্সি স্কেলিং ও নিডল-ইন-আ-হেস্ট্যাক**<br>*লং-কনটেক্সট সিস্টেম — রিং-অ্যাটেনশন, RoPE ফ্রিকোয়েন্সি স্কেলিং ও নিডল-ইন-আ-হেস্ট্যাক* | 304 শব্দ |
| খণ্ড ২০ | **কৃত্রিম প্রজ্ঞার দিগন্ত — মাল্টিমডালিটি, রিজনিং টোকেন ও টেস্ট-টাইম কম্পিউট**<br>*কৃত্রিম প্রজ্ঞার দিগন্ত — মাল্টিমডালিটি, রিজনিং টোকেন ও টেস্ট-টাইম কম্পিউট* | 280 শব্দ |

শাকিল মাহমুদ মনোগ্রাফ সিরিজ • প্রথম খণ্ড

# টোকেনাইজেশন বটলনেক — বাইট-পেয়ার এনকোডিং, ওয়ার্ডপিস ও সাবওয়ার্ড শব্দভাণ্ডার

*টোকেনাইজেশন বটলনেক — বাইট-পেয়ার এনকোডিং, ওয়ার্ডপিস ও সাবওয়ার্ড শব্দভাণ্ডার*

353 শব্দ • November 4, 2026 • ওয়েব সংস্করণ




**চিত্র ১:** *যান্ত্রিক অটোমাটা ও রামোন লুলের কম্বিনেটোরিয়াল আর্স ম্যাগনা এবং গটফ্রিড ভিলহেম লাইবনিজের ক্যারেক্টারিস্টিকা ইউনিভার্সালিস ও ক্যালকুলাস র‍্যাটিওসিনেটরের রূপরেখা।*

একটি আধুনিক লার্জ ল্যাঙ্গুয়েজ মডেল (LLM) তার প্রথম গাণিতিক সমীকরণ সম্পন্ন করার আগেই মানুষের স্বাভাবিক ভাষাকে একটি বিচ্ছিন্ন সংকেতমালায় রূপান্তর করতে হয়: একে বলা হয় **টোকেনাইজেশন (Tokenization)**। এটি কোনো সাধারণ প্রস্তুতিমূলক ধাপ নয়; এটি হলো আধুনিক জেনারেটিভ এআই-এর সবচেয়ে মৌলিক সীমাবদ্ধতা বা বটলনেক। গণিতে ভুল করা, বানানে অক্ষরের সংখ্যা ভুল গণনা করা বা বহুভাষিক মডেলের ব্যয় বৃদ্ধির মতো সমস্যাগুলো মূলত এই সাবওয়ার্ড টোকেনাইজেশনের কারণে ঘটে থাকে।

## ১. শব্দ বনাম অক্ষরের চিরন্তন টানাপোড়েন

স্বাভাবিক ভাষা প্রক্রিয়াকরণে দুটি চরম পন্থা রয়েছে:

- **শব্দ-ভিত্তিক টোকেনাইজেশন:** প্রতিটি পূর্ণাঙ্গ শব্দের জন্য একটি ইউনিক আইডি দিলে সিকোয়েন্স দৈর্ঘ্য ছোট হয়, কিন্তু অভিধানের আকার কোটি ছাড়িয়ে যায়। বাংলা বা ফিনিশের মতো রূপমূলগতভাবে জটিল ভাষায় প্রচুর নতুন শব্দ তৈরি হয় যা মডেলের কাছে অপরিচিত ($langletextunkrangle$) থেকে যায়।
- **অক্ষর-ভিত্তিক টোকেনাইজেশন:** মাত্র ২৫৬টি অক্ষর রাখলে কোনো শব্দই অপরিচিত থাকে না, কিন্তু বাক্যের দৈর্ঘ্য ৪ থেকে ১০ গুণ বেড়ে যায়। ট্রান্সফরমারের অ্যাটেনশন কোয়াড্রেটিক ($O(N^2)$) হারে বাড়ায় এটি কম্পিউটেশনাল বিপর্যয় ডেকে আনে।

## ২. বাইট-পেয়ার এনকোডিং (BPE) অ্যালগরিদম

১৯৯৪ সালের ফিলিপ গেজের ডেটা কম্প্রেশন অ্যালগরিদমকে ২০১৬ সালে রিকো সেনরিখ নিউরাল ট্রান্সলেশনে প্রয়োগ করে জন্ম দেন **BPE**। অ্যালগরিদমটি নিচের ধাপে কাজ করে:

1. প্রাথমিকভাবে সমস্ত মৌলিক অক্ষর নিয়ে শব্দভাণ্ডার $mathcalV_0$ তৈরি করা হয়।
2. সম্পূর্ণ করপাসে সবচেয়ে বেশিবার পাশাপাশি আসা দুটি অক্ষরের জোড়া $(c_1, c_2)$ খুঁজে বের করা হয়।

Shakil Mahmud • Research Monograph

3. এই সবচেয়ে ঘন ঘন আসা জোড়াটিকে যুক্ত করে একটি নতুন সাবওয়ার্ড টোকেন তৈরি করা হয় এবং শব্দভাণ্ডারে যোগ করা হয়।
4. এই প্রক্রিয়া হাজার হাজার বার পুনরাবৃত্তি করে ৩২ হাজার থেকে ১ লাখ ২৮ হাজার আকারের একটি সর্বোত্তম শব্দভাণ্ডার তৈরি হয়।

ঘন ঘন ব্যবহৃত শব্দগুলো একটি একক টোকেনে পরিণত হয়, আর বিরল শব্দগুলো অর্থপূর্ণ ক্ষুদ্রাংশে ভেঙে যায় (যেমন: “antigravity” *to* `['anti', 'gravity']`)।

## ৩. বাইট-লেভেল বিপিই (BBPE)

জিপিটি-২ ও আধুনিক এলএলএমগুলোতে কাঁচা ৮-বিট বাইট (0x00 – 0xFF) থেকে বিপিই শুরু করা হয়। যেহেতু বিশ্বের যেকোনো ভাষা বা কোড আলটিমেটলি UTF-8 বাইট দ্বারা গঠিত, তাই বাইট-লেভেল বিপিই-তে **কখনোই কোনো আননোন টোকেন ত্রুটি হতে পারে না**।

## ৪. টোকেনাইজেশনের পার্শ্বপ্রতিক্রিয়া ও বহুভাষিক বৈষম্য

ইংরেজি ভাষায় প্রতি শব্দে প্রায় ১.৩টি টোকেন খরচ হয়, কিন্তু বাংলা বা হিন্দিতে একটি শব্দের জন্যই ৩ থেকে ৬টি টোকেন খরচ হয়। এর ফলে অ-ইংরেজি ভাষার ব্যবহারকারীদের জন্য এআই ব্যবহারের খরচ বহুগুণ বেড়ে যায় এবং তাদের কনটেক্সট উইন্ডো দ্রুত সংকুচিত হয়ে পড়ে।



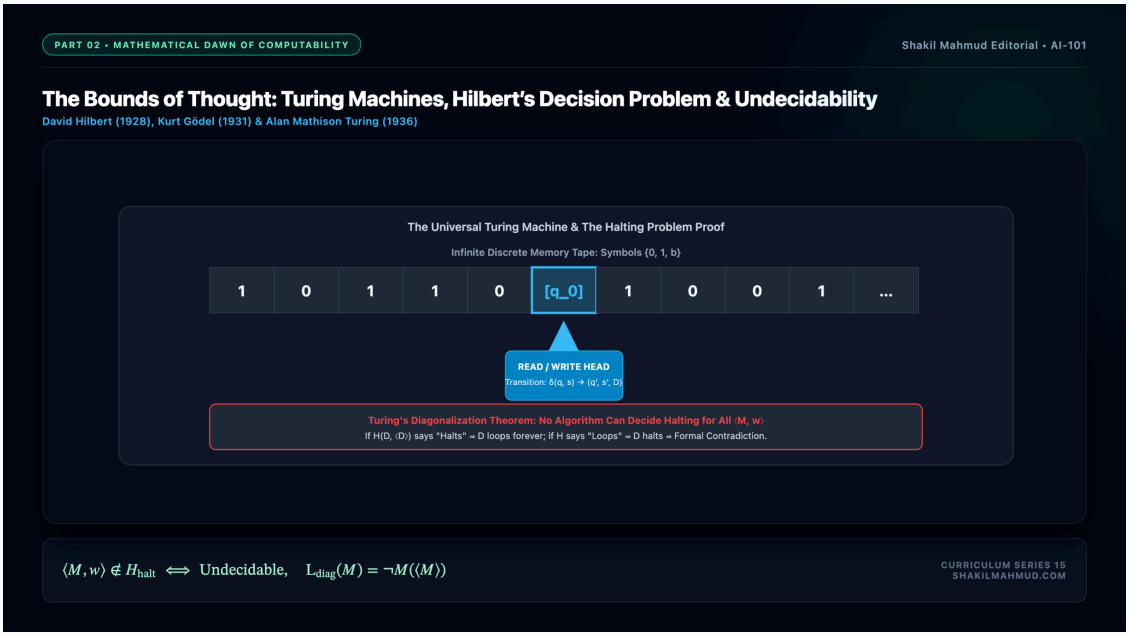
*সাইটেশন: মাহমুদ, শাকিল (2026)। "টোকেনাইজেশন বটলনেক — বাইট-পেয়ার এনকোডিং, ওয়ার্ডপিস ও সাবওয়ার্ড শব্দভাণ্ডার"। লার্জ ল্যাঙ্গুয়েজ মডেল ও জেনারেটিভ এআই, খণ্ড ১।*
*https://shakilmahmud.com/bn/large-language-models-part-01-tokenization-byte-pair-encoding-bpe-subwords-bn/*

শাকিল মাহমুদ মনোগ্রাফ সিরিজ • দ্বিতীয় খণ্ড

# নিউরাল স্কেলিং ল — কাপলান বনাম চিনচিলা কম্পিউট অপ্টিমালিটি

*নিউরাল স্কেলিং ল — কাপলান বনাম চিনচিলা কম্পিউট অপ্টিমালিটি*

326 শব্দ • November 4, 2026 • ওয়েব সংস্করণ




**চিত্র ২:** *ইউনিভার্সাল টুরিং মেশিনের ডিসক্রিট টেপ আর্কিটেকচার এবং হল্টিং প্রবলেমের অনির্ণেয়তা প্রমাণকারী ডায়াগোনালাইজেশন মেকানিজ্ম।*

কম্পিউটার বিজ্ঞানের ইতিহাসে কোনো অ্যালগরিদমের কার্যকারিতা পূর্বে নির্ভর করত মানুষের বুদ্ধিমত্তা ও সূক্ষ্ম কৌশলের ওপর। কিন্তু লার্জ ল্যাঙ্গুয়েজ মডেলের উন্মেষ এই ধারণা আমূল বদলে দিয়েছে। আবিষ্কৃত হয়েছে **নিউরাল স্কেলিং ল (Neural Scaling Laws)**: মডেলের আকার, ডেটাসেটের পরিধি এবং কম্পিউটেশনাল শক্তির সাথে ক্রস-এন্ট্রপি লস কোনো খেয়ালখুশিমতো আচরণ করে না; এটি পদার্থবিজ্ঞানের নিয়মের মতো এক নিখুঁত পাওয়ার-ল (Power-Law) মেনে হ্রাস পায়। আধুনিক এআই-এর কোটি কোটি ডলারের বিনিয়োগ নিয়ন্ত্রিত হয় এই স্কেলিং সূত্রের মাধ্যমে।

## ১. কাপলান স্কেলিং ল (OpenAI, 2020)

২০২০ সালে ওপেনএআই-এর জ্যারেড কাপলান প্রথম বৃহৎ পরিসরে স্কেলিং নিয়ম প্রকাশ করেন। মোট ফ্লপস $Capprox6ND$-এর সাপেক্ষে তারা দেখান যে লস সূচকীয় হারে কমে। তবে কাপলানের দল দাবি করে যে ডেটা বাড়ানোর চেয়ে মডেলের প্যারামিটার সংখ্যা অনেক বেশি দ্রুত বাড়ানো উচিত ( $NproptoC^{0.73}, DproptoC^{0.27}$ )। এর ওপর ভিত্তি করে ১৭৫ বিলিয়ন প্যারামিটারের জিপিটি-৩ তৈরি হয় যা মাত্র ৩০০ বিলিয়ন টোকেনে প্রশিক্ষিত হয়েছিল।

## ২. চিনচিলা বিপ্লব (DeepMind, 2022)

২০২২ সালে গুগল ডিপমাইন্ডের জর্ডান হফম্যান প্রকাশ করেন তাদের ঐতিহাসিক গবেষণা। তারা প্রমাণ করেন যে কাপলানের পরীক্ষায় একটি মৌলিক ভুল ছিল—লার্নিং রেট শিডিউল সঠিকভাবে সমন্বয় না করায় টোকেন বৃদ্ধির সুফল তখন প্রকাশিত হয়নি।

৪০০টিরও বেশি মডেল প্রশিক্ষণের লস বিশ্লেষণ করে তারা ল্যাগ্রাঞ্জিয়ান মাল্টিপ্লায়ার প্রয়োগের মাধ্যমে প্রমাণ করেন:

$$N_{textopt}proptoC^{0.5}, quad D_{textopt}proptoC^{0.5}$$

অর্থাৎ, কম্পিউটেশনাল শক্তি বৃদ্ধির সাথে সাথে **প্যারামিটার এবং ট্রেনিং টোকেনকে হুবহু সমান অনুপাতে (১:১) বাড়াতে হবে**।

Shakil Mahmud • Research Monograph

## ৩. চিনচিলা রুল অব থাম্ব (১ প্যারামিটারে ২০ টোকেন)

হফম্যান দেখান যে সর্বোত্তম কার্যকারিতার জন্য **প্রতি ১টি প্যারামিটারের বিপরীতে কমপক্ষে ২০টি টোকেন** প্রশিক্ষণ দেওয়া বাধ্যতামূলক। ডিপমাইন্ড মাত্র ৭০ বিলিয়ন প্যারামিটারের **Chinchilla** মডেল তৈরি করে যা ১.৪ ট্রিলিয়ন টোকেনে প্রশিক্ষিত হয়ে ১৭৫ বিলিয়নের জিপিটি-৩ এবং ২৮০ বিলিয়নের গোফারকে প্রতিটি বেঞ্চমার্কে পরাজিত করে।

## ৪. ইনফারেন্স-অপ্টিমাল ওভার-ট্রেনিং ও এলএলএএমএ (LLaMA)

চিনচিলা কেবল ট্রেনিং কম্পিউটের হিসাব দেয়। কিন্তু বাস্তবে মডেলকে দীর্ঘদিন কোটি কোটি মানুষের প্রশ্নের উত্তর দিতে হয়। ফলে মেটা তাদের LLaMA সিরিজে চিনচিলার সীমানা ছাড়িয়ে একটি ৭ বিলিয়ন মডেলকে ১ ট্রিলিয়ন টোকেন এবং ৭০ বিলিয়ন মডেলকে ২ ট্রিলিয়নেরও বেশি টোকেনে প্রশিক্ষণ দেয়, যা আধুনিক ওপেন সোর্স এআই বিপ্লবের ভিত্তি স্থাপন করে।

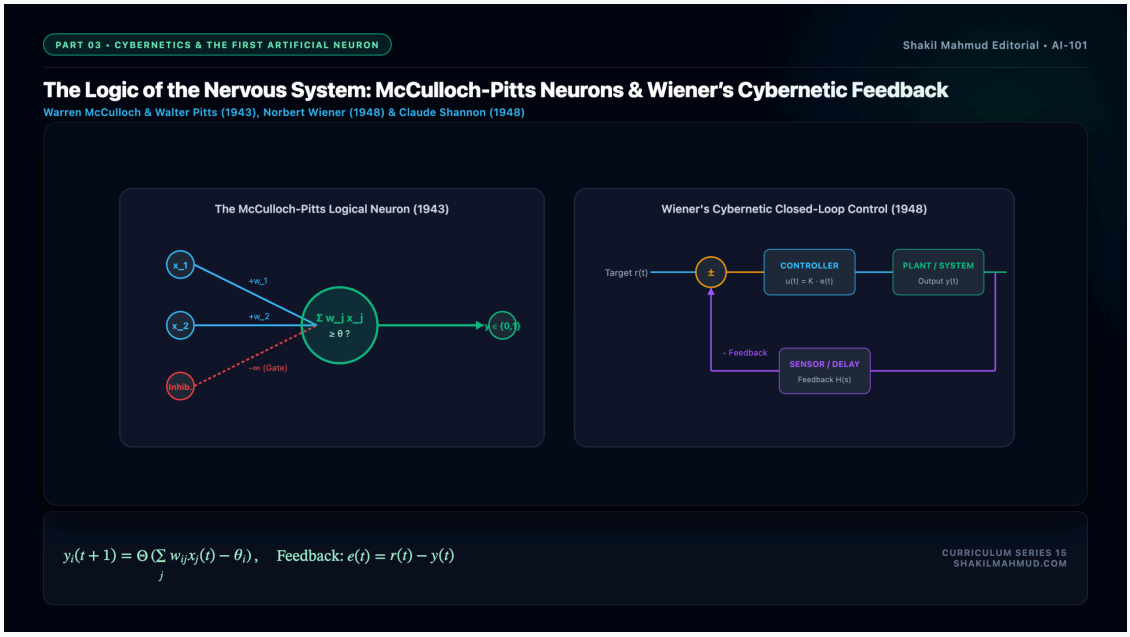
*সাইটেশন: মাহমুদ, শাকিল (2026)। "নিউরাল স্কেলিং ল — কাপলান বনাম চিনচিলা কম্পিউট অপ্টিমালিটি"। লার্জ ল্যাঙ্গুয়েজ মডেল ও জেনারেটিভ এআই, খণ্ড ২।*
*https://shakilmahmud.com/bn/large-language-models-part-02-neural-scaling-laws-kaplan-chinchilla-optimality-bn/*

শাকিল মাহমুদ মনোগ্রাফ সিরিজ • তৃতীয় খণ্ড

# ডেনস ট্রান্সফরমার আর্কিটেকচার — জিপিটি-৩, পালম, এলএলএএমএ ও আধুনিক রেসিপি

*ডেনস ট্রান্সফরমার আর্কিটেকচার — জিপিটি-৩, পালম, এলএলএএমএ ও আধুনিক রেসিপি*

283 শব্দ • November 4, 2026 • ওয়েব সংস্করণ




**চিত্র ৩:** *ম্যাককুলাক-পিটস বাইনারি লজিক্যাল নিউরন এবং নরবার্ট উইনারের সাইবারনেটিক ক্লোজড-লুপ নেগেটিভ ফিডব্যাক নিয়ন্ত্রণ ব্যবস্থা।*

২০১৭ সালের আদি ট্রান্সফরমার এবং ২০২৩ সালের LLaMA ও Mistral-এর মধ্যবর্তী সময়ে অটোরিগ্রেসিভ ল্যাঙ্গুয়েজ মডেলের কাঠামোগত নকশায় এক নীরব বিবর্তন ঘটে গেছে। জিপিটি-১ থেকে জিপিটি-৩ পর্যন্ত মডেলগুলো ভাসওয়ানির আদি স্থাপত্যের খুব কাছাকাছি ছিল। কিন্তু আধুনিক ট্রিলিয়ন-টোকেন ফাউন্ডেশন মডেলগুলো এক অত্যন্ত দক্ষ ও সমন্বিত মানদণ্ডে এসে উপনীত হয়েছে: **Pre-RMSNorm**, **SwiGLU নন-লিনিয়ারিটি**, **Rotary Position Embeddings (RoPE) এবং বায়াসহীন আনটাইড ওয়েট প্রক্ষেপণ**।

## ১. অটোরিগ্রেসিভ ডিকোডার স্তরের আধুনিকায়ন

একটি কজাল ল্যাঙ্গুয়েজ মডেল সম্ভাব্যতা চেইন রুল মেনে পরবর্তী টোকেন প্রেডিক্ট করে। সময় ও মেমরির ব্যবহারে ছয় বছরে চারটি প্রধান পরিবর্তন সার্বজনীন রূপ পেয়েছে:

1. **Post-LN থেকে Pre-LN:** নরমালাইজেশনকে রেসিডুয়াল কানেকশনের ভেতরে নিয়ে আসায় শত শত স্তর গভীর হলেও গ্রেডিয়েন্ট ভ্যানিশ হয় না।
2. **LayerNorm থেকে RMSNorm:** গড় বিয়োগ করার ঝামেলা বাদ দিয়ে কেবল রুট মিন স্কয়ার স্কেলিং করায় গতি ২০% বৃদ্ধি পায়।
3. **ReLU/GELU থেকে SwiGLU:** সাধারণ নন-লিনিয়ার লেয়ারের বদলে গেটেড বাইলিনিয়ার প্রক্ষেপণ মডেলের ধারণক্ষমতা বহুগুণ বাড়ায়।
4. **Absolute Positional Encoding থেকে RoPE:** জটিল সমতলে ভেক্টর ঘোরানোর মাধ্যমে যেকোনো দৈর্ঘ্যের টেক্সট অনুধাবনের সুযোগ সৃষ্টি হয়।

## ২. সুইগ্লু (SwiGLU) অ্যাক্টিভেশন মেকানিজম

২০২০ সালে নোয়াম শাযির **SwiGLU** উদ্ভাবন করেন। এটি সাধারণ ফিড-ফরওয়ার্ড স্তরের মতো একটি ম্যাট্রিক্স গুণ না করে, ইনপুটকে দুটি সমান্তরাল শাখায় বিভক্ত করে একটিকে SiLU অ্যাক্টিভেশন দিয়ে গেট হিসেবে ব্যবহার করে অপরটির সাথে গুণ করে:

Shakil Mahmud • Research Monograph

$$textSwiGLU(x) = (textswish(xW_1)odotxW_2)W_3$$

সমান কম্পিউটেশনাল খরচে সুইগ্লু উল্লেখযোগ্যভাবে উন্নত পারফরম্যান্স দেয়, যা PaLM, LLaMA এবং Mistral-এর অবিচ্ছেদ্য অংশ।

## ৩. আনটাইড আউটপুট হেড (Untied Embeddings)

পূর্বে প্যারামিটার বাঁচাতে ইনপুট এম্বেডিং ম্যাট্রিক্স এবং আউটপুট ক্লাসিফায়ার ম্যাট্রিক্স একই রাখা হতো ($W_{texthead} = W_e^T$)। কিন্তু আধুনিক শব্দভাণ্ডার সুবিশাল (V = 128k – 256k)। গবেষকরা দেখিয়েছেন যে ইনপুটে শব্দের অর্থগত নৈকট্য দরকার হলেও আউটপুটে অসম্ভব শব্দগুলোকে কঠোরভাবে শূন্য করার জন্য তীব্র জ্যামিতিক বিভাজন দরকার। দুটি ম্যাট্রিক্সকে পৃথক (Untied) করে দিলে মডেলের যুক্তিপ্রয়োগ ক্ষমতা বহুগুণ বৃদ্ধি পায়।



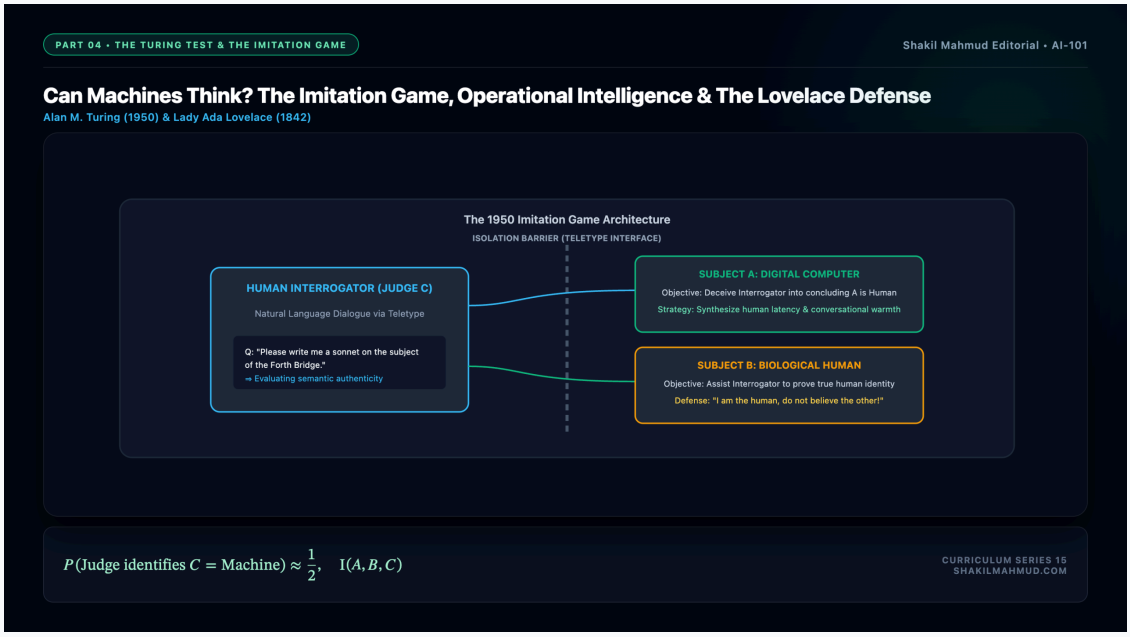
*সাইটেশন: মাহমুদ, শাকিল (2026)। "ডেনস ট্রান্সফরমার আর্কিটেকচার — জিপিটি-৩, পালম, এলএলএএমএ ও আধুনিক রেসিপি"। লার্জ ল্যাঙ্গুয়েজ মডেল ও জেনারেটিভ এআই, খণ্ড ৩।*
*https://shakilmahmud.com/bn/large-language-models-part-03-dense-transformer-architectures-llama-mistral-paradigms-bn/*

শাকিল মাহমুদ মনোগ্রাফ সিরিজ • চতুর্থ খণ্ড

# কেভি ক্যাশ ও ইনফারেন্স লেটেন্সি — পেজড-অ্যাটেনশন, ভিএলএলএম ও সার্ভিং মেমরি

*কেভি ক্যাশ ও ইনফারেন্স লেটেন্সি — পেজড-অ্যাটেনশন, ভিএলএলএম ও সার্ভিং মেমরি*

291 শব্দ • November 4, 2026 • ওয়েব সংস্করণ




**চিত্র ৪:** *অ্যালান টুরিংয়ের ১৯৫০ সালের ইমিটেশন গেম ষ্ট্রাইপার্টাইট স্থাপত্য, যা ব্লাইন্ড টেলিটাইপ কথোপকথনের মাধ্যমে কৃত্রিম বুদ্ধিমত্তা মূল্যায়ন করে।*

একটি লার্জ ল্যাঙ্গুয়েজ মডেলকে প্রশিক্ষণ দেওয়ার সময় মূল চ্যালেঞ্জ থাকে কম্পিউট বা গণনার গতি (Compute-Bound)। কিন্তু যখন সেই মডেল ব্যবহারকারীর প্রশ্নের উত্তর তৈরি করে, তখন পরিস্থিতি সম্পূর্ণ উল্টে যায়: তখন গতি নির্ধারিত হয় মেমরি ব্যান্ডউইডথের মাধ্যমে (Memory-Bandwidth Bound)। প্রতিটি টোকেন তৈরির জন্য জিপিইউকে তার বিলিয়ন বিলিয়ন ওজন প্যারামিটার এবং অতীতের সমস্ত ইতিহাসের মেমরি পড়তে হয়। আধুনিক এলএলএম সার্ভিংয়ের সবচেয়ে বড় যুদ্ধক্ষেত্র হলো এই **কি-ভ্যালু (KV) ক্যাশ** এবং বার্কলের উদ্ভাবিত **PagedAttention (vLLM)** আর্কিটেকচার।

## ১. প্রিফিল বনাম ডিকোড পর্ব

ইনফারেন্স দুটি ধাপে সম্পন্ন হয়:

1. **প্রিফিল (Prefill):** ব্যবহারকারীর সম্পূর্ণ প্রম্পট একসাথে সমান্তরালভাবে পড়ে। এটি কম্পিউট-বাউন্ড এবং জিপিইউর টেনসর কোর পূর্ণ গতিতে ব্যবহার করে।
2. **ডিকোড (Decode):** একটি একটি করে নতুন টোকেন তৈরি করে। প্রতিটি ধাপে নতুন টোকেনটির সাথে অতীতের সমস্ত টোকেনের সম্পর্ক নির্ণয় করতে হয়। অতীতের মানগুলো পুনরায় হিসাব না করার জন্য মেমরিতে জমিয়ে রাখা হয়—যাকে বলে **KV Cache**।

## ২. কেভি ক্যাশের ভয়াবহ মেমরি আকার

একটি ৭০ বিলিয়ন প্যারামিটারের LLaMA মডেলে ৩২ জন ব্যবহারকারী যদি ৪০০০ টোকেন করে টেক্সট জেনারেট করেন, তবে তাদের KV ক্যাশের আকার দাঁড়ায় প্রায় **৩৪৪ গিগাবাইট**! এটি মডেলের মূল ওজনের (১৪০ জিবি) চেয়েও দ্বিগুণের বেশি এবং চারটি ৮০ জিবি A100 জিপিইউর মেমরি একাই গ্রাস করে ফেলে।

Shakil Mahmud • Research Monograph

## ৩. চিরায়ত মেমরির বিশৃঙ্খলা ও অপচয়

আগের সার্ভিং সিস্টেমে প্রতিটি ব্যবহারকারীর জন্য সর্বোচ্চ দৈর্ঘ্যের (যেমন ২০৪৮ টোকেন) মেমরি অগ্রিম বরাদ্দ রাখা হতো। বাস্তব জীবনে অধিকাংশ উত্তর ছোট হওয়ায় **৬০% থেকে ৮০% জিপিইউ মেমরি সম্পূর্ণ অব্যবহৃত** পড়ে থাকত।

## ৪. পেজড-অ্যাটেনশন (PagedAttention) ও vLLM

২০২৩ সালে ইউসি বার্কলের গবেষকরা অপারেটিং সিস্টেমের ভার্চুয়াল মেমরি পেজিং ধারণা প্রয়োগ করে **PagedAttention** উদ্ভাবন করেন। এটি কেভি ক্যাশকে ১৬ বা ৩২ টোকেনের ক্ষুদ্র ফিজিক্যাল ব্লকে ভাগ করে। মেমরিতে ধারাবাহিক জায়গা না থাকলেও ব্লক টেবিলের মাধ্যমে তথ্য যুক্ত থাকে। এর ফলে মেমরি অপচয় ৪%-এর নিচে নেমে আসে এবং একই সার্ভারে ২ থেকে ৪ গুণ বেশি মানুষের অনুরোধ একসঙ্গে প্রসেস করা সম্ভব হয়।

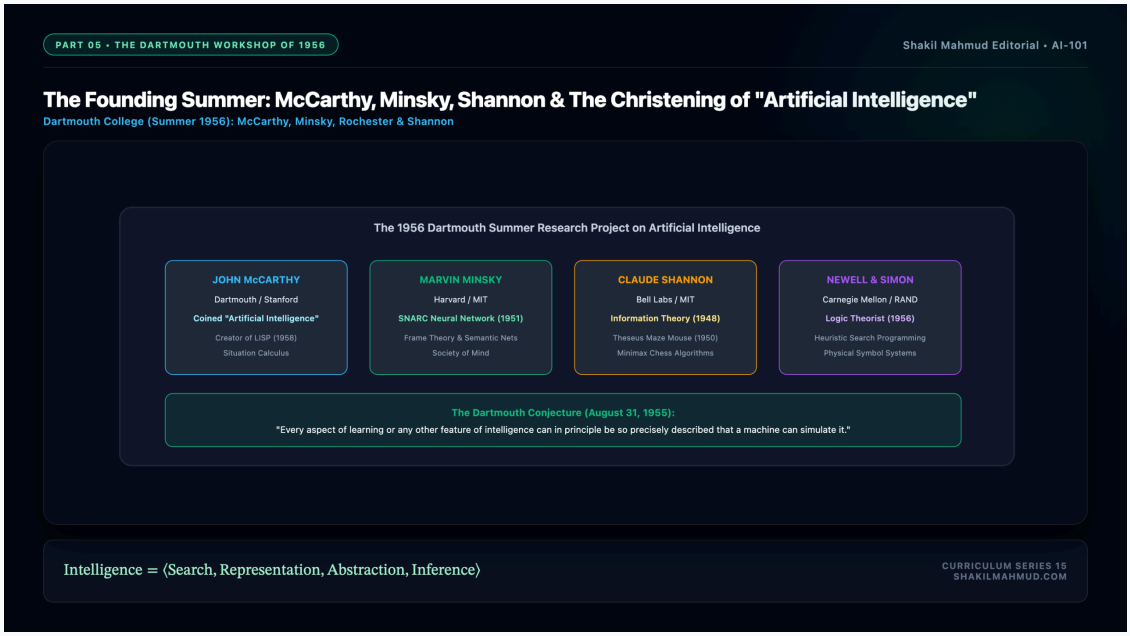
*সাইটেশন: মাহমুদ, শাকিল (2026)। "কেভি ক্যাশ ও ইনফারেন্স লেটেন্সি — পেজড-অ্যাটেনশন, ভিএলএলএম ও সার্ভিং মেমরি"। লার্জ ল্যাঙ্গুয়েজ মডেল ও জেনারেটিভ এআই, খণ্ড ৪।*
*https://shakilmahmud.com/bn/large-language-models-part-04-kv-cache-inference-latency-pagedattention-vllm-bn/*

শাকিল মাহমুদ মনোগ্রাফ সিরিজ • পঞ্চম খণ্ড

# মাল্টি-কোয়েরি ও গ্রুপড-কোয়েরি অ্যাটেনশন — মেমরি ব্যান্ডউইথ সীমানা অতিক্রম

*মাল্টি-কোয়েরি ও গ্রুপড-কোয়েরি অ্যাটেনশন — মেমরি ব্যান্ডউইথ সীমানা অতিক্রম*

441 শব্দ • November 4, 2026 • ওয়েব সংস্করণ




**চিত্র ৫:** *১৯৫৬ সালের ডার্টমাউথ এআই সম্মেলন: জন ম্যাককার্থি, মার্ভিন মিনস্কি, ক্লড শ্যানন এবং নিউয়েল-সাইমনের যুগান্তকারী প্রতিষ্ঠা।*

অটোরিগ্রেসিভ টোকেন তৈরির সময় লার্জ ল্যাঙ্গুয়েজ মডেলগুলো কাঁচা ফ্লোটিং-পয়েন্ট কম্পিউট দ্বারা আটকে থাকে না; বরং তাদের প্রধান বাধা হয়ে দাঁড়ায় **মেমরি ব্যান্ডউইথ (Memory Bandwidth)**। প্রতি একক টোকেন জেনারেট করতে পুরো মডেলের প্যারামিটার এবং বিগত টোকেনগুলোর দীর্ঘ Key-Value টেনসর হিস্ট্রি হাই-ব্যান্ডউইথ মেমরি (HBM) থেকে চিপের ক্যাশে রিড করতে হয়। **মাল্টি-কোয়েরি অ্যাটেনশন (MQA)** এবং **গ্রুপড-কোয়েরি অ্যাটেনশন (GQA)** হলো যুগান্তকারী আর্কিটেকচারাল উদ্ভাবন যা মডেলের যৌক্তিক ক্ষমতা অক্ষুণ্ণ রেখে কেভি-ক্যাশের মেমরি চাপ ৮ গুণ পর্যন্ত কমিয়ে দেয়।

## ১. মেমরি দেয়াল ও অ্যারিথমেটিক ইনটেনসিটি

নিউরাল ইনফারেন্সের কর্মক্ষমতা নির্ধারিত হয় **Arithmetic Intensity** দ্বারা—অর্থাৎ প্রতি বাইট মেমরি স্থানান্তরের বিপরীতে কতগুলো ফ্লোটিং-পয়েন্ট অপারেশন (FLOPs) সম্পন্ন হচ্ছে:

$$mathcal{I} = frac{text{FLOPs}}{text{Bytes Transferred}}$$

চিরায়ত মাল্টি-হেড অ্যাটেনশনে (MHA), যদি হিডেন ডাইমেনশন $d$, কোয়েরি হেড সংখ্যা $H_Q$ এবং সিকোয়েন্স দৈর্ঘ্য $S$ হয়, তবে প্রতিটি হেড তার নিজস্ব Key ও Value প্রজেকশন তৈরি করে:

$$mathbf{K}_h = mathbf{X} mathbf{W}_K^{(h)}, quad mathbf{V}_h = mathbf{X} mathbf{W}_V^{(h)} quad (h = 1, dots, H_Q)$$

প্রতিটি নতুন টোকেন ডিকোড করার সময় পুরো পূর্ববর্তী KV ক্যাশ ($2 times H_Q times S times d_k$ বাইট) মেমরি থেকে টেনে আনতে হয়। FLOPs এবং মেমরি ট্রাফিক উভয়ই $O(S)$ অনুপাতে বাড়ায় অ্যারিথমেটিক ইনটেনসিটি ১ FLOP/বাইট-এর নিচে নেমে আসে। ফলে NVIDIA H100 জিপিইউর মতো শক্তিশালী প্রসেসরের টেনসর কোরগুলো ৯০% সময় ডেটার অপেক্ষায় নিষ্ক্রিয় বসে থাকে।

Shakil Mahmud • Research Monograph

## ২. মাল্টি-কোয়েরি অ্যাটেনশন (MQA): চরম কেভি ভাগাভাগি

২০১৯ সালে নোয়াম শাজির **Multi-Query Attention (MQA)** প্রস্তাব করেন। এখানে Query হেডগুলো স্বাধীন থাকলেও ($H_Q$ টি), সমস্ত কোয়েরি হেড মাত্র একটি অভিন্ন Key এবং Value হেড ($H_{KV} = 1$) ভাগ করে নেয়:

$mathbfQinmathbbR^{BtimesH_Qtimes1timesd_k}, quadmathbfKinmathbbR^{Btimes1timesStimesd_k}, quadmathbfVinmathbbR^{Btimes1timesStimesd_k}$

এর ফলে মেমরি স্থানান্তরের পরিমাণ $H_Q$ গুণ (সাধারণত ৩২ থেকে ৬৪ গুণ) কমে যায়। তবে সমস্ত হেড একই তথ্যের ওপর নির্ভরশীল হওয়ায় জটিল যুক্তি ও দীর্ঘ অনুচ্ছেদ বিশ্লেষণে মডেলের গুণগত মান কিছুটা হ্রাস পায়।

## ৩. গ্রুপড-কোয়েরি অ্যাটেনশন (GQA): সুষম সমাধান

MHA-এর উচ্চ গুণমান এবং MQA-এর ক্ষিপ্র গতির মেলবন্ধন ঘটাতে ২০২৩ সালে অ্যানসলি প্রমুখ উদ্ভাবন করেন **Grouped-Query Attention (GQA)**। এটি LLaMA-2/3 এবং Mistral মডেলে গৃহীত মানদণ্ড। এখানে $H_Q$ সংখ্যক কোয়েরি হেডকে $G$ সংখ্যক গ্রুপে ভাগ করা হয়। প্রতিটি গ্রুপ একটি করে Key-Value হেড ব্যবহার করে ($H_{KV} = G$):

- যখন $G = H_Q$: এটি অবিকল সাধারণ মাল্টি-হেড অ্যাটেনশন (MHA)।
- যখন $G = 1$: এটি পুরোপুরি মাল্টি-কোয়েরি অ্যাটেনশন (MQA)।
- যখন 1 < G < H_Q (সাধারণত $G = 8$): এটি মেমরি ট্রাফিক ৮ গুণ কমিয়ে দেয় কিন্তু মডেলের নিখুঁত বোঝার ক্ষমতা ৯৯% অক্ষুণ্ণ রাখে।

## ৪. গ্রুপড-কোয়েরি সফটম্যাক্স সূত্র

হেড $h$-এর গ্রুপ সূচক $g(h) = lfloorh/(H_Q/G)rfloor$ হলে অ্যাটেনশন সমীকরণ দাঁড়ায়:

$mathbfO_h = textSoftmaxleft(fracmathbfQ_hmathbfK_{g(h)}^Tsqrtd_kright)mathbfV_{g(h)}$

এর ফলে সার্ভারে সমান্তরাল অনুরোধ প্রসেস করার গতি ৩ গুণ থেকে ৫ গুণ পর্যন্ত বৃদ্ধি পায়।

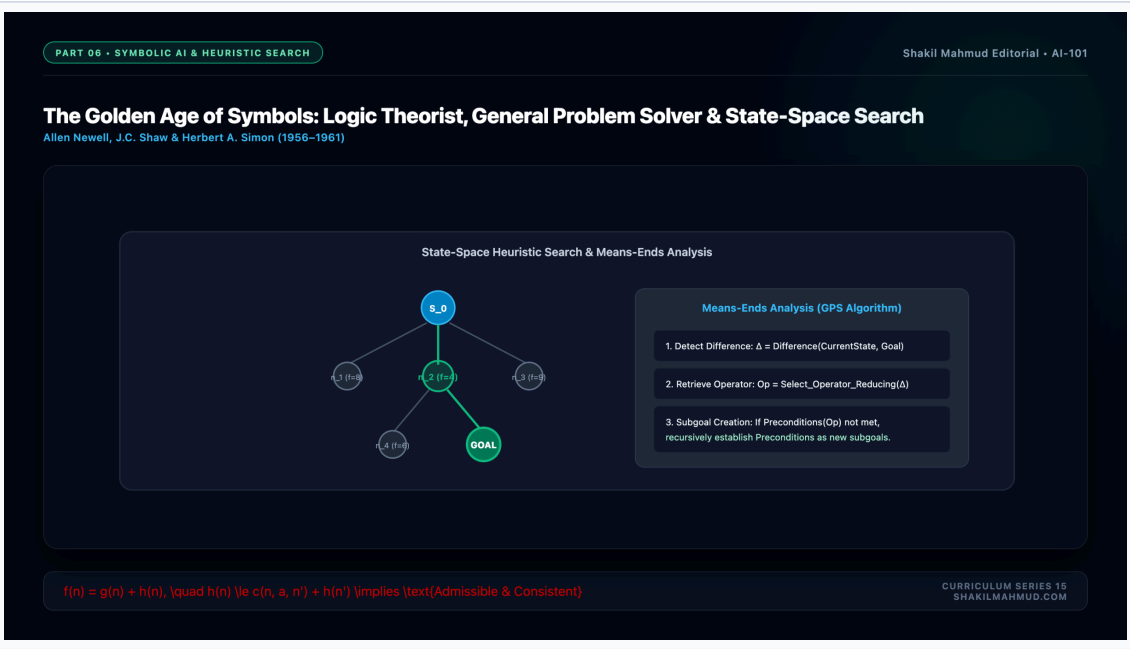
*সাইটেশন: মাহমুদ, শাকিল (2026)। "মাল্টি-কোয়েরি ও গ্রুপড-কোয়েরি অ্যাটেনশন — মেমরি ব্যান্ডউইথ সীমানা অতিক্রম"। লার্জ ল্যাঙ্গুয়েজ মডেল ও জেনারেটিভ এআই, খণ্ড ৫।*
*https://shakilmahmud.com/bn/large-language-models-part-05-multi-query-grouped-query-attention-mqa-gqa-bn/*

শাকিল মাহমুদ মনোগ্রাফ সিরিজ • ষষ্ঠ খণ্ড

# ফ্ল্যাশ-অ্যাটেনশন ও আই/ও সচেতনতা — জিপিইউ এস-র‍্যাম টাইলিং ও অনলাইন সফটম্যাক্স

*ফ্ল্যাশ-অ্যাটেনশন ও আই/ও সচেতনতা — জিপিইউ এস-র‍্যাম টাইলিং ও অনলাইন সফটম্যাক্স*

398 শব্দ • November 4, 2026 • ওয়েব সংস্করণ




**চিত্র ৬:** *স্টেট-স্পেস হিউরিস্টিক সার্চ ট্রি এবং জেনারেল প্রবলেম সলভারের (GPS) মিন্স-এন্ডস অ্যানালাইসিস অ্যালগরিদম।*

ট্রান্সফরমার উদ্ভাবনের পর পাঁচ বছর ধরে সেলফ-অ্যাটেনশনকে নিছক একটি গাণিতিক সমীকরণ মনে করা হতো: $textSoftmax(mathbfQmathbfK^T/sqrtd)mathbfV$। কিন্তু বাস্তবে দীর্ঘ টেক্সটের ক্ষেত্রে অ্যাটেনশন ধীরগতির হয়ে পড়েছিল গাণিতিক জটিলতার ( $O(N^2)$ FLOPs) কারণে নয়, বরং **মেমরি ইনপুট/আউটপুট (IO)** ট্রাফিকের চাপে। ২০২২ সালে ট্রাই দাও এবং ক্রিস্টোফার রে উদ্ভাবন করেন **FlashAttention**, যা হার্ডওয়্যারের অভ্যন্তরীণ মেমরি বিন্যাসকে সরাসরি বিবেচনায় নিয়ে অ্যাটেনশন গণনায় কোনো প্রকার আনুমানিক ভুল ছাড়াই ২ থেকে ৪ গুণ গতি এবং মেমরি খরচ ১০ গুণ পর্যন্ত কমিয়ে আনে।

## ১. জিপিইউর অভ্যন্তরীণ মেমরি স্তর: SRAM বনাম HBM

একটি আধুনিক জিপিইউতে (যেমন NVIDIA A100 বা H100) মূলত দুটি মেমরি স্তর থাকে:

- **হাই ব্যান্ডউইথ মেমরি (HBM / DRAM):** বিশাল ধারনক্ষমতা (৮০ জিবি), কিন্তু তুলনামূলকভাবে ধীর গতি (১.৫–৩.৩৫ টেরাবাইট/সেকেন্ড) এবং উচ্চ লেটেন্সি।
- **স্ট্যাটিক র‍্যাম (SRAM / Shared Memory):** অত্যন্ত ক্ষিপ্র গতি (প্রায় ১৯ টেরাবাইট/সেকেন্ড), কিন্তু আকার খুবই ক্ষুদ্র (প্রতি প্রসেসরে মাত্র ১৯২ কিলোবাইট)।

চিরায়ত পাইটর্চে অ্যাটেনশন হিসাব করার সময় প্রতি ধাপে মধ্যবর্তী বিশাল টেনসরগুলোকে ধীরগতির HBM-এ লিখতে ও পড়তে হয়:

1. HBM থেকে $mathbfQ, mathbfK$ এনে গুণ করে $mathbfS = mathbfQmathbfK^Tinmathbb{R}^{NtimesN}$ তৈরি করা হয় $rightarrow$ আবার HBM-এ লেখা হয়।
2. HBM থেকে $mathbfS$ পড়ে সফটম্যাক্স করে $mathbfP$ বের করা হয় $rightarrow$ আবার HBM-এ লেখা হয়।
3. HBM থেকে $mathbfP$ ও $mathbfV$ পড়ে গুণ করে চূড়ান্ত আউটপুট $mathbfO$ বের করা হয়।

Shakil Mahmud • Research Monograph

সিকোয়েন্স দৈর্ঘ্য ৩২,০০০ হলে এই $NtimesN$ ম্যাট্রিক্সের আকার দাঁড়ায় **প্রতি ব্যাচ আইটেমে ২ গিগাবাইট**! ফলে জিপিইউর কম্পিউট কোরগুলো অলস বসে থেকে কেবল ডেটা ট্রাফিকের শিকার হয়।

## ২. অনলাইন সফটম্যাক্স (Online Softmax) কৌশল

অ্যাটেনশনকে ক্ষুদ্র ব্লকে ভাগ করার প্রধান অন্তরায় ছিল সফটম্যাক্সের হর: $sum_{j=1}^{N} exp(S_{ij})$। পুরো সারির সমস্ত সংখ্যা না জানা পর্যন্ত সাধারণ সফটম্যাক্স হিসাব করা যায় না।

FlashAttention এই সমস্যার সমাধান করে **অনলাইন সফটম্যাক্স** সূত্রের সাহায্যে। ধরা যাক দুটি ব্লকের মান $x^{(1)}$ ও $x^{(2)}$, যাদের লোকাল সর্বোচ্চ মান $m_1$ ও $m_2$। সম্মিলিত সর্বোচ্চ মান হবে:

$m_{textnew} = max(m_1, m_2)$

মধ্যবর্তী এক্সপোনেনশিয়াল যোগফলকে স্কেলিং গুণক দিয়ে সাথে সাথে আপডেট করা যায়:

l_{text{new}} = l_1 e^{m_1 – m_{text{new}}} + l_2 e^{m_2 – m_{text{new}}}

একইভাবে সঞ্চিত আউটপুট ভেক্টর $mathbf{O}$-কেও ধাপে ধাপে আপডেট করা হয়:

mathbf{O}_{text{new}} = mathbf{O}_1 left(frac{l_1 e^{m_1 – m_{text{new}}}}{l_{text{new}}}right) + mathbf{O}_2 left(frac{e^{m_2 – m_{text{new}}}}{l_{text{new}}}right)

## ৩. টাইলিং অ্যালগরিদম ও ফলাফল

অনলাইন সফটম্যাক্সের কল্যাণে FlashAttention ইনপুট ম্যাট্রিক্সগুলোকে ক্ষুদ্র ব্লকে বিভক্ত করে যা সরাসরি দ্রুতগতির SRAM-এর ভেতরে এঁটে যায়। $NtimesN$ ম্যাট্রিক্সটি কখনোই ধীরগতির HBM মেমরিতে লেখা হয় না। মেমরি ট্রাফিক $O(N^2)$ থেকে কমে $O(Nd)$ তে রূপান্তরিত হয়, যার ফলে দীর্ঘ প্রম্পট ও বৃহৎ এলএলএম দ্রুততম গতিতে প্রশিক্ষণ ও ইনফারেন্স পরিচালনা করতে পারে।

*সাইটেশন: মাহমুদ, শাকিল (2026)। "ফ্ল্যাশ-অ্যাটেনশন ও আই/ও সচেতনতা — জিপিইউ এস-র‍্যাম টাইলিং ও অনলাইন সফটম্যাক্স"। লার্জ ল্যাঙ্গুয়েজ মডেল ও জেনারেটিভ এআই, খণ্ড ৬।*
*https://shakilmahmud.com/bn/large-language-models-part-06-flashattention-gpu-sram-tiling-online-softmax-bn/*

শাকিল মাহমুদ মনোগ্রাফ সিরিজ • সপ্তম খণ্ড

# মডেল কোয়ান্টাইজেশন তত্ত্ব — FP16, INT8, INT4, GPTQ, AWQ ও আউটলায়ার ডাইমেনশন

*মডেল কোয়ান্টাইজেশন তত্ত্ব — FP16, INT8, INT4, GPTQ, AWQ ও আউটলায়ার ডাইমেনশন*

325 শব্দ • November 4, 2026 • ওয়েব সংস্করণ




**চিত্র ৭:** *ফ্র্যাঙ্ক রোজেনব্ল্যাটের মার্ক ১ পারসেপট্রন রেটিনা, মোটর পটেনশিওমিটার ওজন এবং নোভিকফের কনভার্জেন্স হাইপারপ্লেন।*

৭০ বিলিয়ন প্যারামিটারের একটি লার্জ ল্যাঙ্গুয়েজ মডেল FP16 প্রিসিশনে চালাতে কেবল ওজন সংরক্ষণ করতেই ১৪০ গিগাবাইট গ্রাফিক্স মেমরি (VRAM) প্রয়োজন হয়। **মডেল কোয়ান্টাইজেশন (Quantization)** অবিচ্ছিন্ন ১৬-বিট ফ্লোটিং-পয়েন্ট ওজনকে ৮-বিট, ৪-বিট কিংবা ২-বিট ইন্টিজারে সংকুচিত করে। এটি কেবল সংখ্যা কেটে ছোট করা নয়; বরং বহুমাত্রিক ওজন ম্যাট্রিক্সকে এমনভাবে ক্ষুদ্র গ্রিডে প্রক্ষেপণ করা যাতে মডেলের মূল যুক্তি ও নির্ভুলতা অক্ষুণ্ণ থাকে।

## ১. কোয়ান্টাইজেশনের প্রাথমিক গণিত

ইউনিফর্ম কোয়ান্টাইজেশন বাস্তব সংখ্যার বিস্তারকে $[x_{min}, x_{max}]$ নির্দিষ্ট বিট $b$-এর পূর্ণসংখ্যার গ্রিডে রূপান্তর করে:

$q = textcliple ft(le ftl floor fracxsrightrceil + z, -2^{b-1}, 2^{b-1} - 1right)$

যেখানে $s$ হলো স্কেল ফ্যাক্টর এবং $z$ হলো জিরো-পয়েন্ট অফসেট:

s = frac{x_{max} – x_{min}}{2^b – 1}, quad z = leftlfloor -frac{x_{min}}{s} rightrceil – 2^{b-1}

ডিকোয়ান্টাইজেশনের মাধ্যমে পুনরায় ফ্লোটিং-পয়েন্ট সংখ্যা অনুমান করা হয়: hat{x} = s(q – z)।

## ২. স্কেলিংয়ে আউটলায়ার ডাইমেনশনের আবির্ভাব

মডেলের আকার যখন ৬.৭ বিলিয়ন প্যারামিটার অতিক্রম করে, তখন এর অ্যাক্টিভেশনে এক বিস্ময়কর ঘটনা ঘটে: কিছু নির্দিষ্ট চ্যানেলের মান স্বাভাবিকের চেয়ে ১০০ গুণ বেশি বৃদ্ধি পায় (|X| > 100)। সাধারণ কোয়ান্টাইজেশন ব্যবহার করলে এই বিশালাকার আউটলায়ারগুলোর কারণে স্কেল ফ্যাক্টর বড় হয়ে যায় এবং সাধারণ ক্ষুদ্র মানগুলো শূন্যে বিলীন হয়ে মডেল নষ্ট হয়ে যায়। **LLM.int8()** এই সমস্যার সমাধান করে: ০.১% আউটলায়ার চ্যানেলকে নিখুঁত FP16-এ গুণ করা হয় এবং বাকি ৯৯.৯% চ্যানেলকে দ্রুতগতির INT8-এ প্রসেস করা হয়।

Shakil Mahmud • Research Monograph

## ৩. অপটিমাল ব্রেন সার্জন ও GPTQ

ওজনগুলোকে বিচ্ছিন্নভাবে গোলগাল করার বদলে **GPTQ** প্রতিটি স্তরের আউটপুট ত্রুটি কমানোর অপ্টিমাইজেশন পরিচালনা করে:

argmin_{widehat{mathbf{W}}} |mathbf{W}mathbf{X} – widehat{mathbf{W}}mathbf{X}|_2^2

টেইলর সিরিজের দ্বিতীয় অর্ডার হেসিয়ান ম্যাট্রিক্স $mathbf H^{-1} = (2mathbf X mathbf X^T + lambda mathbf I)^{-1}$ ব্যবহার করে একটি ওজনকে ৪-বিটে নামানোর ফলে যে ত্রুটি তৈরি হয়, বাকি ওজনগুলোকে সমন্বয় করে সেই ত্রুটি তাৎক্ষণিকভাবে সংশোধন করে দেওয়া হয়:

Delta mathbf{w} = – frac{w_q – hat{w}_q}{[mathbf{H}^{-1}]_{qq}} mathbf{H}^{-1}_{:, q}

## ৪. অ্যাক্টিভেশন-সচেতন কোয়ান্টাইজেশন (AWQ)

AWQ পদ্ধতি প্রমাণ করে যে সমস্ত ওজন সমান গুরুত্বপূর্ণ নয়। যে ওজনগুলোর সাথে বড় অ্যাক্টিভেশন গুণ হয় (শীর্ষ ১%), কেবল সেগুলোকে সুরক্ষিত রাখলে ৪-বিট কোয়ান্টাইজেশনেও মডেলের কোনো গুণগত ক্ষতি হয় না। এর ফলে একটি সাধারণ গেমিং জিপিইউতেও ৭০ বিলিয়ন প্যারামিটারের বিশ্বমানের মডেল স্বাচ্ছন্দ্যে চালানো সম্ভব হয়।

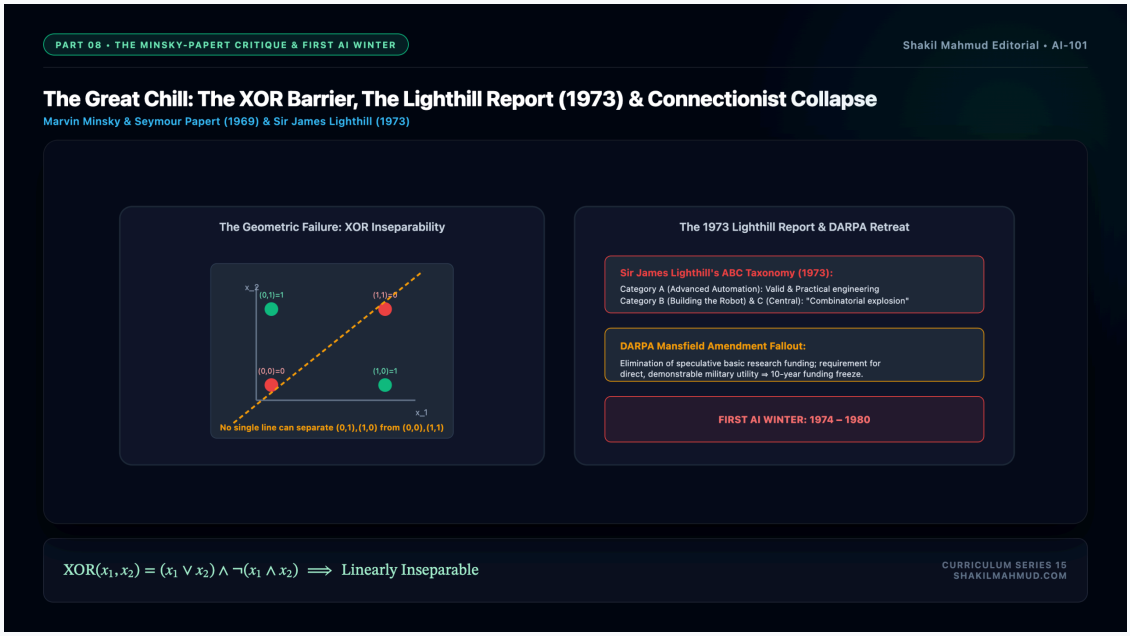
*সাইটেশন: মাহমুদ, শাকিল (2026)। "মডেল কোয়ান্টাইজেশন তত্ত্ব — FP16, INT8, INT4, GPTQ, AWQ ও আউটলায়ার ডাইমেনশন"। লার্জ ল্যাঙ্গুয়েজ মডেল ও জেনারেটিভ এআই, খণ্ড ৭। https://shakilmahmud.com/bn/large-language-models-part-07-model-quantization-fp16-int8-int4-gptq-awq-bn/*

শাকিল মাহমুদ মনোগ্রাফ সিরিজ • অষ্টম খণ্ড

# প্যারামিটার-দক্ষ ফাইন-টিউনিং — লোরা, কিউ-লোরা ও ইন্ট্রিনসিক র‍্যাংকের জ্যামিতি

*প্যারামিটার-দক্ষ ফাইন-টিউনিং — লোরা, কিউ-লোরা ও ইন্ট্রিনসিক র‍্যাংকের জ্যামিতি*

365 শব্দ • November 4, 2026 • ওয়েব সংস্করণ




**চিত্র ৮:** *মিনস্কি ও প্যাপার্টের প্রমাণিত এক্সওআর (XOR) লিনিয়ার ইনসেপারেবিলিটি এবং ১৯৭৩ সালের লাইটহিল রিপোর্ট পরবর্তী প্রথম এআই শীত।*

৭০ বিলিয়ন প্যারামিটারের একটি ফাউন্ডেশন মডেলকে সম্পূর্ণ ফাইন-টিউনিং করতে সবকটি ওজন আপডেট এবং অ্যাডাম অপটিমাইজারের মোমেন্টাম সংরক্ষণ করতে ৫৬০ গিগাবাইটেরও বেশি গ্রাফিক্স মেমরি (VRAM) প্রয়োজন হয়। **প্যারামিটার-এফিশিয়েন্ট ফাইন-টিউনিং (PEFT)** এই সমস্যার যুগান্তকারী সমাধান দিয়েছে। **ইন্ট্রিনসিক র‍্যাংক হাইপোথিসিস**-এর ওপর ভিত্তি করে **LoRA (Low-Rank Adaptation)** এবং **QLoRA** মূল মডেলের ওজনকে স্থির রেখে ক্ষুদ্র ম্যাট্রিক্স যোগ করে একটিমাত্র সাধারণ জিপিইউতেই বিশাল মডেল ফাইন-টিউনিংয়ের সুযোগ করে দেয়।

## ১. ইন্ট্রিনসিক ডাইমেনশন হাইপোথিসিস

গবেষণায় দেখা গেছে যে একটি বৃহৎ নিউরাল নেটওয়ার্কের ওজন ম্যাট্রিক্স $mathbfWinmathbbR^{dtimesk}$ কোটি কোটি ডাইমেনশনে অবস্থান করলেও, কোনো নির্দিষ্ট নতুন কাজের জন্য প্রয়োজনীয় পরিবর্তন $DeltamathbfW$ একটি **অত্যন্ত নিম্ন-মাত্রিক সাবস্পেসের (Low Rank $rlld$)** মধ্যে সীমাবদ্ধ থাকে। ফলে শত কোটি প্যারামিটার পরিবর্তন না করে সামান্য কিছু প্যারামিটার টিউন করলেই মডেলটি চমৎকারভাবে নতুন কাজ শিখে ফেলে।

## ২. লোরার গাণিতিক কাঠামো

২০২১ সালে এডওয়ার্ড হু প্রমুখ LoRA উদ্ভাবন করেন। এখানে ওজন আপডেট $DeltamathbfW$-কে দুটি ক্ষুদ্র ম্যাট্রিক্স $mathbfB$ এবং $mathbfA$-এর গুণফল হিসেবে প্রকাশ করা হয়:

mathbf{W}’ = mathbf{W}_0 + Delta mathbf{W} = mathbf{W}_0 + frac{alpha}{r}(mathbf{B}mathbf{A})

যেখানে:

- $mathbfW_0$: মূল মডেলের স্থির বা ফ্রোজেন ওজন।
- $mathbfBinmathbbR^{dtimesr}$ এবং $mathbfAinmathbbR^{rtimesk}$: প্রশিক্ষণযোগ্য ক্ষুদ্র ম্যাট্রিক্স, যেখানে র‍্যাংক $r$ সাধারণত ৪, ৮, বা ১৬।
- $alpha$: স্কেলিং প্যারামিটার।

Shakil Mahmud • Research Monograph

শুরুতে $mathbf{A}$-কে গাউসিয়ান বিন্যাস দ্বারা এবং $mathbf{B}$-কে সম্পূর্ণ শূন্য দ্বারা ইনিশিয়ালাইজ করা হয়। ফলে প্রশিক্ষণের প্রথম মুহূর্তে $Deltamathbf{W} = mathbf{B}mathbf{A} = mathbf{0}$, যার অর্থ মডেলটির আদি আচরণ একদম অপরিবর্তিত থাকে।

## ৩. ব্যাকপ্রোপাগেশন ও জিরো-ইনফারেন্স ওভারহেড

ফরোয়ার্ড পাসে ইনপুট $mathbf{x}$ উভয় শাখায় প্রবাহিত হয়:

$$mathbf{h} = mathbf{W}_0mathbf{x} + frac{alpha}{r}mathbf{B}mathbf{A}mathbf{x}$$

ব্যাকপ্রোপাগেশনের সময় কেবল ক্ষুদ্র ম্যাট্রিক্স $mathbf{A}$ ও $mathbf{B}$-এর গ্রাডিয়েন্ট বের করা হয়। মূল মডেলের জন্য কোনো অপটিমাইজার স্টেট জমা রাখতে হয় না, যার ফলে মেমরি খরচ ৯৯% পর্যন্ত হ্রাস পায়। ফাইন-টিউনিং শেষে অ্যাডাপ্টারের ওজনকে মূল ওজনের সাথে যোগ করে দেওয়া যায় ($mathbf{W}_{textfinal} = mathbf{W}_0 + frac{alpha}{r}mathbf{B}mathbf{A}$), যার ফলে ইনফারেন্সের সময় কোনো অতিরিক্ত সময় বা লেটেন্সি তৈরি হয় না।

## ৪. কিউ-লোরা (QLoRA): ৪-বিট কোয়ান্টাইজড ফাইন-টিউনিং

২০২৩ সালে টিম ডেটমার্স উদ্ভাবন করেন **QLoRA**। এটি মূল মডেলটিকে অত্যন্ত নিখুঁত ৪-বিট **NormalFloat4 (NF4)** ফরম্যাটে সংকুচিত করে রাখে এবং তার ওপর ১৬-বিট LoRA অ্যাডাপ্টার যুক্ত করে। ডাবল কোয়ান্টাইজেশন ও পেজড অপটিমাইজারের সমন্বয়ে এটি একটিমাত্র সাধারণ জিপিইউতে ৬৫ বিলিয়ন মডেল সম্পূর্ণ নিখুঁতভাবে ফাইন-টিউন করার বৈপ্লবিক ক্ষমতা প্রদান করে।



*সাইটেশন: মাহমুদ, শাকিল (2026)। "প্যারামিটার-দক্ষ ফাইন-টিউনিং — লোরা, কিউ-লোরা ও ইন্ট্রিনসিক র‍্যাংকের জ্যামিতি"। লার্জ ল্যাঙ্গুয়েজ মডেল ও জেনারেটিভ এআই, খণ্ড ৮।*
*https://shakilmahmud.com/bn/large-language-models-part-08-parameter-efficient-fine-tuning-lora-qlora-intrinsic-rank-bn/*

শাকিল মাহমুদ মনোগ্রাফ সিরিজ • নবম খণ্ড

# স্পার্স মিক্সচার-অব-এক্সপার্টস (MoE) — রাউটার গেট, লোড ব্যালান্সিং ও সুইচ ট্রান্সফরমার

*স্পার্স মিক্সচার-অব-এক্সপার্টস (MoE) — রাউটার গেট, লোড ব্যালান্সিং ও সুইচ ট্রান্সফরমার*

364 শব্দ • November 4, 2026 • ওয়েব সংস্করণ




**চিত্র ৯:** *ক্লাসিক্যাল এক্সপার্ট সিস্টেমের গঠন: নলেজ বেস রুলস, ব্যাকওয়ার্ড চেইনিং ইনফারেন্স ইঞ্জিন এবং সার্টেইনটি ফ্যাক্টর ক্যালকুলাস।*

একটি সাধারণ ডেন্স ল্যাঙ্গুয়েজ মডেলে প্যারামিটার সংখ্যা বাড়ানোর অর্থ হলো প্রতি টোকেন প্রক্রিয়াকরণে সমপরিমাণ কম্পিউটেশনাল ব্যয় বেড়ে যাওয়া। **স্পার্স মিক্সচার-অব-এক্সপার্টস (MoE)** প্যারামিটার সংখ্যা ও কম্পিউটেশনের এই বাধ্যতামূলক সংযোগকে সম্পূর্ণ ছিন্ন করে দেয়। পুরো নেটওয়ার্ককে ক্ষুদ্র ক্ষুদ্র বিশেষায়িত সাব-নেটওয়ার্ক বা “এক্সপার্ট”-এ বিভক্ত করে এবং প্রতিটি টোকেনকে মাত্র ১ বা ২টি সেরা এক্সপার্টের কাছে পাঠিয়ে MoE আর্কিটেকচার (যেমন Mixtral 8x7B বা GPT-4) কম খরচে শত শত বিলিয়ন প্যারামিটারের শক্তি উন্মোচন করে।

## ১. কন্ডিশনাল কম্পিউটেশনের দর্শন

চিরায়ত ট্রান্সফরমারে প্রতিটি টোকেনকে হুবহু একই ফিড-ফরোয়ার্ড নেটওয়ার্ক (FFN) দিয়ে যেতে হয়—তা একটি ব্যাকরণগত দাঁড়ি-কমা হোক কিংবা জটিল কোয়ান্টাম সমীকরণ। MoE পদ্ধতিতে প্রতিটি টোকেনের ধরন অনুযায়ী নেটওয়ার্কের একটি ক্ষুদ্র অংশ সক্রিয় হয়:

$$mathbf{y} = sum_{i=1}^{E} G(mathbf{x})_i E_i(mathbf{x})$$

যেখানে $E$ হলো মোট বিশেষজ্ঞ বা এক্সপার্ট নেটওয়ার্ক সংখ্যা এবং $G(mathbf{x})$ হলো স্পার্স রাউটার গেটের সম্ভাব্যতা ভেক্টর।

## ২. গেটিং মেকানিজম ও টপ-$k$ রাউটিং

২০১৭ সালে নোয়াম শাজির স্পার্স টপ-$k$ গেটিং মেকানিজম প্রস্তাব করেন। ইনপুট টোকেন $mathbf{x}$-কে একটি রাউটিং ম্যাট্রিক্স $mathbf{W}_g$ দিয়ে গুণ করে প্রতিটি এক্সপার্টের স্কোর নির্ণয় করা হয়:

$$H(mathbf{x}) = mathbf{x}mathbf{W}_g$$

এরপর কেবল শীর্ষ $k$ (সাধারণত $k = 2$) সংখ্যক বিশেষজ্ঞের স্কোরকে অক্ষুণ্ণ রেখে বাকি সব মানকে $-infty$ করে সফটম্যাক্স করা হয়:

$$G(mathbf{x}) = textSoftmax(textTopK(H(mathbf{x}), k))$$

Shakil Mahmud • Research Monograph

বাকি সমস্ত এক্সপার্টের মান শূন্য হওয়ায় তাদের জন্য কোনো হিসাব করতে হয় না। যেমন Mixtral 8x7B মডেলে মোট ৪৭ বিলিয়ন প্যারামিটার থাকা সত্ত্বেও প্রতিটি টোকেন কেবল ১৩ বিলিয়ন প্যারামিটার ব্যবহার করে, ফলে মডেলটি অবিশ্বাস্য গতিতে কাজ করে।

## ৩. রাউটার কলাপ্স (Router Collapse) ও এক্সপার্ট ধারণক্ষমতা

নিয়ন্ত্রণহীনভাবে প্রশিক্ষণ দিলে একটি ভয়াবহ সংকট দেখা দেয় যাকে বলে **রাউটার কলাপ্স**। শুরুতে সামান্য ভালো পারফর্ম করা দু-একটি এক্সপার্টের কাছে রাউটার সমস্ত টোকেন পাঠাতে শুরু করে, ফলে বাকি এক্সপার্টরা কিছুই শিখতে পারে না। একপর্যায়ে পুরো বিশাল মডেলটি মাত্র দু-একটি সাধারণ মডেলে পতিত হয়। এছাড়া প্রতি জিপিইউতে এক্সপার্টের জন্য নির্দিষ্ট ধারণক্ষমতা (Capacity Limit) থাকে; মাত্রাতিরিক্ত টোকেন আসলে অতিরিক্ত টোকেনগুলো ড্রপ বা বাতিল হয়ে যায়।

## ৪. অক্সিলিয়ারি লোড ব্যালান্সিং লস

রাউটার কলাপ্স রোধ করতে এবং সমস্ত বিশেষজ্ঞের মধ্যে টোকেনের সুষম বণ্টন নিশ্চিত করতে একটি সহায়ক **Load Balancing Loss** যুক্ত করা হয়:

$$mathcalL_{textaux} = alphacdotEsum_{i=1}^{E} f_i P_i$$

যেখানে $f_i$ হলো এক্সপার্ট $i$-তে প্রেরিত বাস্তব টোকেনের অনুপাত এবং $P_i$ হলো রাউটারের সামগ্রিক সম্ভাব্যতা। এই গাণিতিক শর্ত রাউটারকে বাধ্য করে প্রতিটি এক্সপার্টকে সমানভাবে ব্যবহার করতে, যার ফলে হার্ডওয়্যারের প্রতিটি জিপিইউ সমান দক্ষতায় কাজ করে।

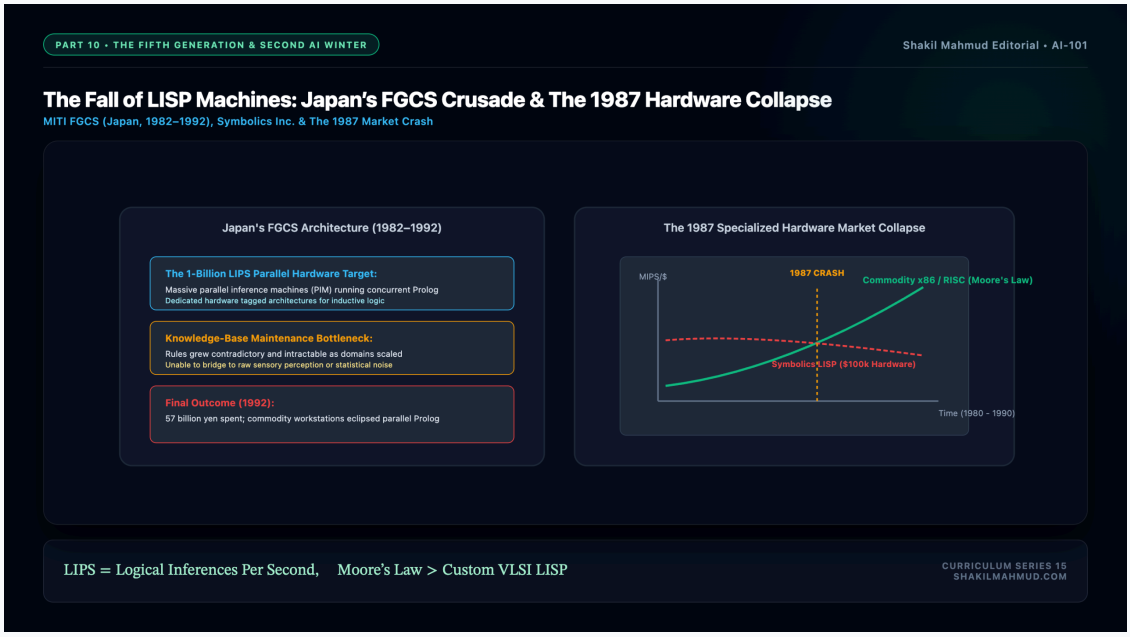
*সাইটেশন: মাহমুদ, শাকিল (2026)। "স্পার্স মিক্সচার-অব-এক্সপার্টস (MoE) — রাউটার গেট, লোড ব্যালান্সিং ও সুইচ ট্রান্সফরমার"। লার্জ ল্যাঙ্গুয়েজ মডেল ও জেনারেটিভ এআই, খণ্ড ৯।*
*https://shakilmahmud.com/bn/large-language-models-part-09-sparse-mixture-of-experts-moe-router-switch-bn/*

শাকিল মাহমুদ মনোগ্রাফ সিরিজ • দশম খণ্ড

# প্রি-ট্রেনিং গতিবিদ্যা ও ডেটা ইঞ্জিনিয়ারিং — মিনহাশ ডিডুপ্লিকেশন ও সিন্থেটিক ডেটাবেস

*প্রি-ট্রেনিং গতিবিদ্যা ও ডেটা ইঞ্জিনিয়ারিং — মিনহাশ ডিডুপ্লিকেশন ও সিন্থেটিক ডেটাবেস*

310 শব্দ • November 4, 2026 • ওয়েব সংস্করণ




**চিত্র ১০:** *জাপানের পঞ্চম প্রজন্মের প্যারালাল প্রোলগ প্রকল্প এবং ১৯৮৭ সালের লিস্প (LISP) মেশিন ও বিশেষায়িত হার্ডওয়্যার বাজারের আকস্মিক পতন।*

আধুনিক জেনারেটিভ এআইয়ের মূল দর্শন হলো: *মডেলের আর্কিটেকচার গৌণ; ডেটার মানই হলো মূল অ্যালগরিদম*। একটি স্টেট-অব-দ্য-আর্ট ফাউন্ডেশন মডেল প্রি-ট্রেনিং করতে ১০ থেকে ১৫ ট্রিলিয়ন টোকেন প্রয়োজন হয়। ইন্টারনেটের কাঁচা ডেটাসেট (Common Crawl)-এর ৮৫%-ই অপ্রয়োজনীয় আবর্জনা, স্প্যাম ও নিম্নমানের লেখায় ভরা। আধুনিক লার্জ ল্যাঙ্গুয়েজ মডেলগুলোর মেধা নির্ভর করে পেটাবাইট স্কেলের উন্নত **ডেটা ইঞ্জিনিয়ারিং পাইপলাইনের** ওপর, যা মিনহাশ হ্যাশিং, পারপ্লেক্সিটি ফিল্টারিং ও সিন্থেটিক ডেটার সমন্বয়ে গঠিত।

## ১. সদৃশ ডকুমেন্ট অপসারণ: MinHash LSH

ওয়েব স্ক্র্যাপিংয়ের ফলে লাখ লাখ ডুপ্লিকেট ও কপি-পেস্ট পেজ জমা হয়। হুবহু হ্যাশ মিলিয়ে এগুলো ধরা যায় না কারণ ডকুমেন্টের ভেতর সামান্য সময় বা বিজ্ঞাপন আলাদা থাকে। **MinHash Locality-Sensitive Hashing (LSH)** অত্যন্ত ক্ষিপ্র গতিতে প্রায়-অভিন্ন ডকুমেন্ট চিহ্নিত করে।

একটি ডকুমেন্টের $k$-শিঙ্গল সেটের মধ্যবর্তী সাদৃশ্যকে **জ্যাকার্ড সিমিলারিটি (Jaccard Similarity)** দ্বারা সংজ্ঞায়িত করা হয়:

$$J(D_1, D_2) = frac|S(D_1) cap S(D_2)||S(D_1) cup S(D_2)|$$

এলোমেলো পারমিউটেশন $pi$ প্রয়োগ করলে দুটি ডকুমেন্টের ক্ষুদ্রতম হ্যাশ মান মিলে যাওয়ার সম্ভাব্যতা হুবহু তাদের জ্যাকার্ড সিমিলারিটির সমান:

$$Pr[min(pi(S(D_1))) = min(pi(S(D_2)))] = J(D_1, D_2)$$

১২৮টি স্বাধীন হ্যাশ ফাংশনকে কয়েকটি ব্যান্ডে ভাগ করে ট্রিলিয়ন টোকেন থেকে কোটি কোটি অপ্রয়োজনীয় ডুপ্লিকেট দূর করা হয়। ডুপ্লিকেট দূর না করলে মডেল নির্দিষ্ট তথ্য মুখস্থ করে ফেলে এবং সাধারণীকরণের ক্ষমতা হারিয়ে ফেলে।

## ২. পারপ্লেক্সিটি ও হিউরিস্টিক ফিল্টারিং

ডিডুপ্লিকেশনের পর ডেটাকে কঠোর মান যাচাইয়ের ভেতর দিয়ে যেতে হয়:

Shakil Mahmud • Research Monograph

- **হিউরিস্টিক নিয়মাবলী:** টেক্সটে যতিচিহ্নের অনুপাত, স্টপ-ওয়ার্ডের উপস্থিতি এবং লাইনের দৈর্ঘ্য যাচাই করে স্প্যাম দূর করা হয়।
- **ভাষা শনাক্তকরণ:** ফাস্টটেক্সট ক্লাসিফায়ার দিয়ে অনাকাঙ্ক্ষিত মিশ্র ভাষার পেজ বাদ দেওয়া হয়।
- **পারপ্লেক্সিটি ফিল্টারিং:** উইকিপিডিয়া বা উচ্চমানের জার্নালের ওপর প্রশিক্ষিত একটি ক্ষুদ্র ল্যাঙ্গুয়েজ মডেল দিয়ে ডকুমেন্টের পারপ্লেক্সিটি মাপা হয়। ভাষা অতিরিক্ত এলোমেলো ও অসংলগ্ন হলে তা সরাসরি বর্জন করা হয়।

## ৩. সিন্থেটিক ডেটা ও টেক্সটবুক বিপ্লব

ইন্টারনেটের প্রাকৃতিক মানসম্মত লেখার ভাণ্ডার ফুরিয়ে আসায় আধুনিক প্রি-ট্রেনিংয়ে **সিন্থেটিক ডেটা (Synthetic Data)** অপরিহার্য হয়ে উঠেছে। মাইক্রোসফটের *Phi* সিরিজের উদ্ভাবন প্রমাণ করেছে যে উচ্চমানের শিক্ষক মডেল দিয়ে তৈরি নিখুঁত পাঠ্যপুস্তকধর্মী ডেটাসেট ব্যবহার করলে অনেক ছোট মডেলও বিশালাকার মডেলের সমান জ্ঞান ও যুক্তি প্রদর্শন করতে পারে।

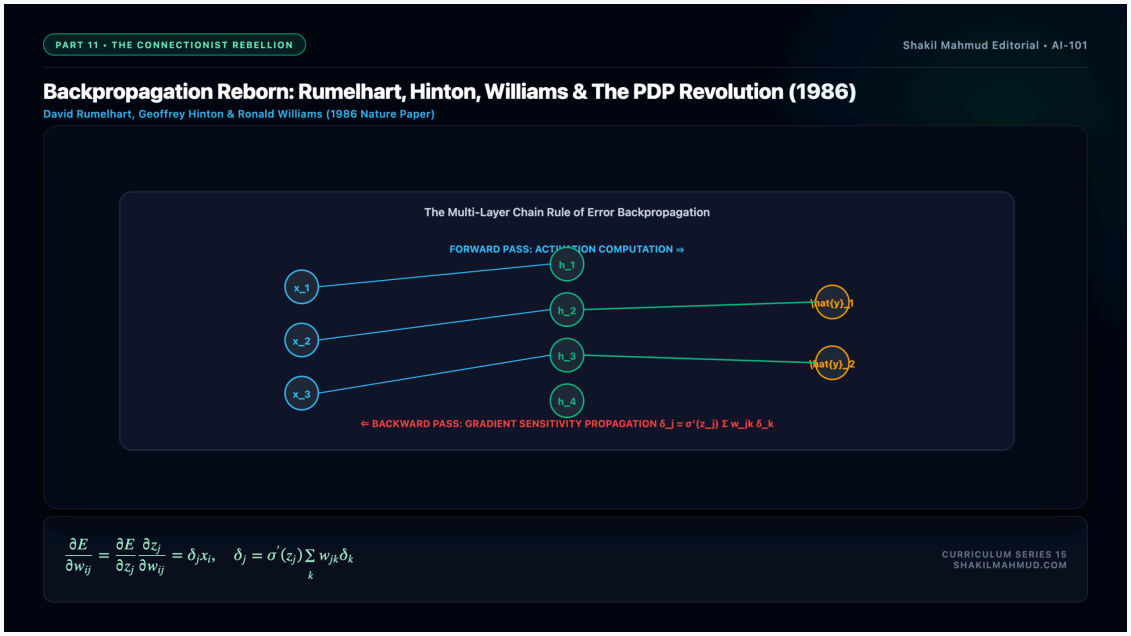
*সাইটেশন: মাহমুদ, শাকিল (2026)। "প্রি-ট্রেনিং গতিবিদ্যা ও ডেটা ইঞ্জিনিয়ারিং — মিনহাশ ডিডুপ্লিকেশন ও সিন্থেটিক ডেটাবেস"। লার্জ ল্যাঙ্গুয়েজ মডেল ও জেনারেটিভ এআই, খণ্ড ১০।*
*https://shakilmahmud.com/bn/large-language-models-part-10-pretraining-dynamics-data-engineering-minhash-synthetic-bn/*

শাকিল মাহমুদ মনোগ্রাফ সিরিজ • একাদশ খণ্ড

# সুপারভাইজড ফাইন-টিউনিং ও ইন্সট্রাকশন টিউনিং — ক্রস-এনট্রপি গতিবিদ্যা

*সুপারভাইজড ফাইন-টিউনিং ও ইন্সট্রাকশন টিউনিং — ক্রস-এনট্রপি গতিবিদ্যা*

255 শব্দ • November 4, 2026 • ওয়েব সংস্করণ




**চিত্র ১১:** *মাল্টি-লেয়ার ব্যাকপ্রোপাগেশন অ্যালগরিদম: ফরওয়ার্ড অ্যাক্টিভেশন পাস এবং চেইন রুলের সাহায্যে ব্যাকওয়ার্ড এরর সংবেদনশীলতা বিস্তার।*

একটি সাধারণ প্রি-ট্রেইনড বেস মডেল শুধুই পরবর্তী টোকেন অনুমানের পরিসংখ্যানিক ইঞ্জিন। তাকে যদি জিজ্ঞেস করা হয় *“বাংলাদেশের রাজধানী কোথায়?”*, তবে সে হয়তো উত্তর না দিয়ে পরবর্তী প্রশ্ন হিসেবে লিখতে পারে *“ভারতের রাজধানী কোথায়? পাকিস্তানের রাজধানী কোথায়?”* কারণ ইন্টারনেটে প্রশ্নমালার টেক্সট সেভাবেই থাকে। **সুপারভাইজড ফাইন-টিউনিং (SFT)** এবং **ইন্সট্রাকশন টিউনিং** এই অন্ধ টেক্সট কমপ্লিশনকে একজন বুদ্ধিমান, সুশৃঙ্খল ও নির্দেশনাকারী সহকারীতে রূপান্তর করে।

## ১. চ্যাট মার্কআপ ও কাঠামোগত নিরাপত্তা (ChatML)

ব্যবহারকারীর নির্দেশ এবং সিস্টেমের কঠোর নিয়মাবলী আলাদা রাখতে আধুনিক এলএলএমগুলো **ChatML** বা বিশেষ টোকেন কাঠামো ব্যবহার করে:

```
<|im_start|>system
আপনি একজন সহায়ক শিক্ষক।<|im_end|>
<|im_start|>user
দ্বিঘাত সমীকরণ কী?<|im_end|>
<|im_start|>assistant
যে সমীকরণে অজ্ঞাত রাশির সর্বোচ্চ ঘাত দুই...<|im_end|>
```

এই বিশেষ টোকেনগুলো ব্যবহারকারী টাইপ করতে পারে না, ফলে প্রম্পট ইনজেকশন বা হ্যাকিং থেকে মডেল সুরক্ষিত থাকে।

## ২. প্রম্পট-রেসপন্স লস মাস্কিং

ফাইন-টিউনিংয়ের সময় একটি বড় ভুল হলো পুরো টেক্সটের ওপর লস হিসাব করা। ব্যবহারকারী কী প্রম্পট লিখেছে তা মুখস্থ করার জন্য মডেলকে গ্রেড দেওয়া উচিত নয়; মডেলকে কেবল তার দেওয়া উত্তরের নির্ভুলতার জন্য অপটিমাইজ করতে হবে।

যদি প্রম্পট $১dots T$ টোকেন এবং উত্তর $T + ১dots N$ টোকেন হয়, তবে লস মাস্কিং সমীকরণ দাঁড়ায়:

Shakil Mahmud • Research Monograph

mathcal{L}_{text{SFT}}(theta) = – sum_{t=T+1}^N log P_theta(x_t mid x_1, dots, x_{t-1})

ব্যবহারকারীর প্রম্পটের জন্য লস শূন্য করে দেওয়া হয়, যার ফলে মডেলের সমস্ত গ্রেডিয়েন্ট আপডেট কেবল সঠিক উত্তরের শৈলী ও তথ্য গঠনে ব্যয় হয়।

## ৩. মাল্টি-টাস্ক সাধারণীকরণ

গুগলের Flan গবেষণায় দেখা গেছে যে শত শত বৈচিত্র্যময় কাজের নির্দেশ (সারাংশ তৈরি, অনুবাদ, কোডিং, গণিত) একসাথে ফাইন-টিউন করলে মডেলের অভ্যন্তরীণ বোধশক্তি নাটকীয়ভাবে বৃদ্ধি পায়। ফলে জীবনে কখনো দেখেনি এমন সম্পূর্ণ নতুন কোনো নির্দেশ পেলেও মডেলটি নিখুঁতভাবে তার সমাধান বের করতে সক্ষম হয়।

*সাইটেশন: মাহমুদ, শাকিল (2026)। "সুপারভাইজড ফাইন-টিউনিং ও ইন্সট্রাকশন টিউনিং — ক্রস-এনট্রপি গতিবিদ্যা"। লার্জ ল্যাঙ্গুয়েজ মডেল ও জেনারেটিভ এআই, খণ্ড ১১।*
*https://shakilmahmud.com/bn/large-language-models-part-11-supervised-fine-tuning-multi-task-instruction-tuning-bn/*

শাকিল মাহমুদ মনোগ্রাফ সিরিজ • দ্বাদশ খণ্ড

# রিইনফোর্সমেন্ট লার্নিং ফ্রম হিউম্যান ফিডব্যাক (RLHF) — ব্র্যাডলি-টেরি রিওয়ার্ড মডেল ও পিপিও

*রিইনফোর্সমেন্ট লার্নিং ফ্রম হিউম্যান ফিডব্যাক (RLHF) — ব্র্যাডলি-টেরি রিওয়ার্ড মডেল ও পিপিও*

314 শব্দ • November 4, 2026 • ওয়েব সংস্করণ




**চিত্র ১২:** *ইয়ান লেকানের লেনাট-৫ (LeNet-5) কনভোলিউশনাল নেটওয়ার্ক: লোকাল রিসেপ্টিভ ফিল্ড, ওজন শেয়ারিং এবং সাবস্যাম্পলিং স্তরবিন্যাস।*

সুপারভাইজড ফাইন-টিউনিং একটি মডেলকে উত্তর লেখার বিন্যাস শেখাতে পারলেও কোনটি বেশি বুদ্ধিদীপ্ত, তথ্যবহুল বা নিরাপদ তা নিখুঁতভাবে সংজ্ঞায়িত করতে পারে না। মানুষের পক্ষে কোটি কোটি নিখুঁত উত্তর নিজ হাতে লিখে দেওয়া অসম্ভব, কিন্তু দুটি উত্তরের মধ্যে কোনটি বেশি ভালো তা বেছে নেওয়া খুবই সহজ। **রিইনফোর্সমেন্ট লার্নিং ফ্রম হিউম্যান ফিডব্যাক (RLHF)** মানুষের এই তুলনামূলক পছন্দকে কাজে লাগিয়ে একটি রিওয়ার্ড মডেল তৈরি করে এবং প্রক্সিমাল পলিসি অপটিমাইজেশনের (PPO) মাধ্যমে মডেলটিকে মানব সমাজের মূল্যবোধ ও নৈতিকতার সাথে সুসংগত করে তোলে।

## ১. ব্র্যাডলি-টেরি প্রেফারেন্স মডেল

একটি প্রম্পট $x$-এর জন্য দুটি মডেলের উত্তর $(y_w, y_l)$-এর মধ্যে মানুষ যদি $y_w$-কে বিজয়ী হিসেবে বেছে নেয়, তবে সেই পছন্দকে ১৯৫২ সালের ক্লাসিক্যাল **ব্র্যাডলি-টেরি সমীকরণ** দ্বারা প্রকাশ করা হয়:

P(y_w succ y_l mid x) = sigma(r_psi(x, y_w) – r_psi(x, y_l)) = frac{1}{1 + e^{-(r_psi(x, y_w) – r_psi(x, y_l))}}

যেখানে $r_psi(x, y)$ হলো একটি স্কেলার রিওয়ার্ড মডেল। মানব পরীক্ষকদের হাজার হাজার পছন্দ-অপছন্দের জোড়া থেকে এই রিওয়ার্ড মডেলটি শিখে ফেলে মানুষ কোন ধরনের উত্তর পছন্দ করে।

## ২. রিওয়ার্ড হ্যাকিং ও কেএল ডাইভারজেন্স পেনাল্টি

রিওয়ার্ড মডেলের ওপর নির্ভর করে মডেলটিকে মুক্তভাবে রিইনফোর্সমেন্ট লার্নিং শেখাতে গেলে একটি নতুন সংকট তৈরি হয়: **রিওয়ার্ড হ্যাকিং**। মডেলটি রিওয়ার্ড ফাংশনের কোনো ফাঁকফোকর খুঁজে নিয়ে অর্থহীন চাটুকারিতামূলক কথা বলে কৃত্রিমভাবে বেশি নম্বর নেওয়ার চেষ্টা করে।

মডেলকে সঠিক ভাষার সীমানায় ধরে রাখতে আদি রেফারেন্স মডেল $pi_{textref}$-এর সাথে **KL Divergence Penalty** যুক্ত করা হয়:

R(x, y) = r_psi(x, y) – beta log frac{pi_theta(y mid x)}{pi_{text{ref}}(y mid x)}

Shakil Mahmud • Research Monograph

এই সমীকরণ মডেলটিকে বেশি নম্বর পাওয়ার পাশাপাশি মূল ভাষার সাবলীল ব্যাকরণ ও ভাব বজায় রাখতে বাধ্য করে।

## ৩. প্রক্সিমাল পলিসি অপটিমাইজেশন (PPO)

ভাষার প্রতিটি শব্দ তৈরি করাকে একটি মার্কভ ডিসিশন প্রসেস (MDP) হিসেবে বিবেচনা করা হয়। নীতি পরিবর্তনের চরম আকস্মিকতা রোধ করতে **PPO** অ্যালগরিদম সম্ভাবনা অনুপাতকে ক্লিপ করে রাখে:

mathcal{L}_{text{PPO}}(theta) = mathbb{E}_t left[ minleft( rho_t(theta) hat{A}_t, , text{clip}(rho_t(theta), 1-epsilon, 1+epsilon) hat{A}_t right) right]

একসাথে চারটি নিউরাল নেটওয়ার্কের (Actor, Critic, Reference, এবং Reward) জটিল সমন্বয়ে RLHF প্রক্রিয়া আধুনিক ChatGPT এবং সমসাময়িক সুশৃঙ্খল এআই সিস্টেমগুলোর ভিত্তি স্থাপন করেছে।

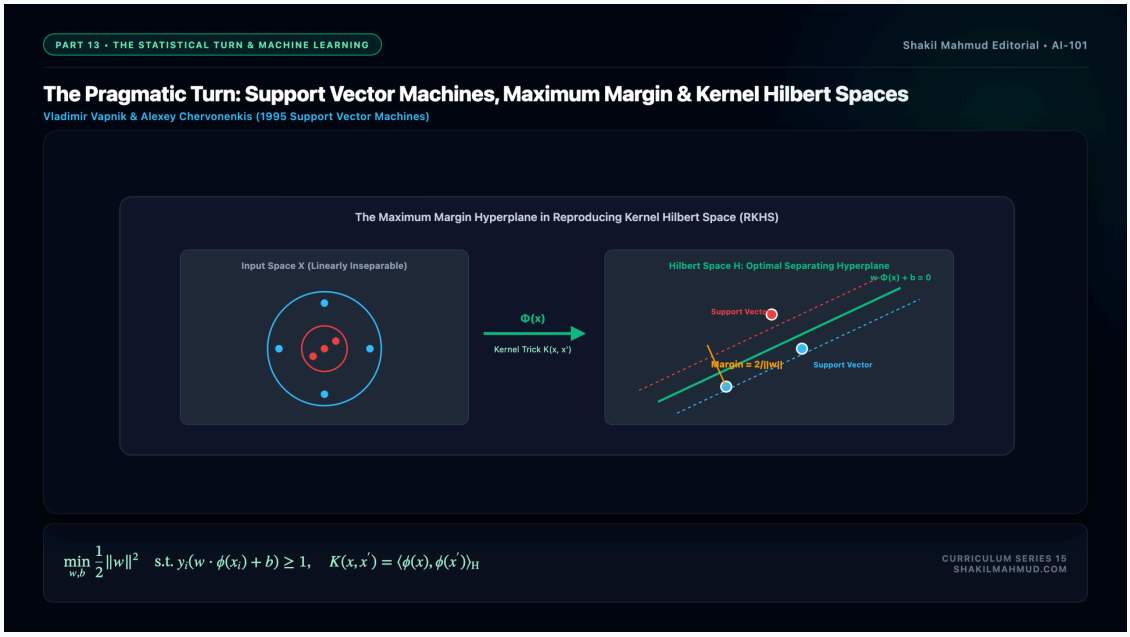
*সাইটেশন: মাহমুদ, শাকিল (2026)। "রিইনফোর্সমেন্ট লার্নিং ফ্রম হিউম্যান ফিডব্যাক (RLHF) — ব্র্যাডলি-টেরি রিওয়ার্ড মডেল ও পিপিও"। লার্জ ল্যাঙ্গুয়েজ মডেল ও জেনারেটিভ এআই, খণ্ড ১২।*
*https://shakilmahmud.com/bn/large-language-models-part-12-rlhf-reward-modeling-bradley-terry-ppo-bn/*

শাকিল মাহমুদ মনোগ্রাফ সিরিজ • ত্রয়োদশ খণ্ড

# ডাইরেক্ট প্রেফারেন্স অপটিমাইজেশন (DPO) — রিওয়ার্ড মডেলবিহীন ইমপ্লিসিট লগ-অডস

*ডাইরেক্ট প্রেফারেন্স অপটিমাইজেশন (DPO) — রিওয়ার্ড মডেলবিহীন ইমপ্লিসিট লগ-অডস*

294 শব্দ • November 4, 2026 • ওয়েব সংস্করণ




**চিত্র ১৩:** *রিপ্রোডিউসিং কার্নেল হিলবার্ট স্পেস সাপোর্ট ভেক্টর মেশিন: ম্যাক্সিমাম মার্জিন হাইপারপ্লেন এবং নন-লিনিয়ার কার্নেল ট্রিক।*

PPO-ভিত্তিক রিইনফোর্সমেন্ট লার্নিং ফ্রম হিউম্যান ফিডব্যাক (RLHF) যুগান্তকারী সাফল্য আনলেও এটি পরিচালনাগতভাবে অত্যন্ত জটিল ও অস্থিতিশীল। একসাথে চারটি বিশাল নিউরাল নেটওয়ার্ক চালানো বিপুল মেমরি ও সূক্ষ্ম হাইপারপ্যারামিটার সামঞ্জস্য দাবি করে। ২০২৩ সালে স্ট্যানফোর্ড বিশ্ববিদ্যালয়ের রাফায়েল রাফাইলোভ প্রমুখ উদ্ভাবন করেন **Direct Preference Optimization (DPO)**। তারা গাণিতিকভাবে প্রমাণ করেন যে কোনো পৃথক রিওয়ার্ড মডেল তৈরি না করে এবং কোনো আরএল স্যাম্পলিং ছাড়াই সরাসরি মানবীয় পছন্দ অনুযায়ী মডেল অপটিমাইজ করা সম্ভব।

## ১. আদর্শ নীতির গাণিতিক সমাধান

চিরায়ত আরএলএইচএফের কেএল-পেনাল্টিযুক্ত লক্ষ্যমাত্রাকে গাণিতিকভাবে বিশ্লেষণ করলে দেখা যায় যে এর সর্বোচ্চ সমাধানের একটি সুনির্দিষ্ট বদ্ধ রূপ (Closed-Form Solution) রয়েছে:

pi^*(y mid x) = frac{1}{Z(x)} pi_{text{ref}}(y mid x) expleft( frac{1}{beta} r(x, y) right)

যেখানে $Z(x)$ হলো একটি পার্টিশন ফাংশন। এটি প্রমাণ করে যে রিওয়ার্ড ফাংশন $r(x, y)$ এবং ভাষা মডেলের সম্ভাব্যতা অনুপাতের মধ্যে একটি প্রত্যক্ষ গাণিতিক সম্পর্ক বিদ্যমান।

## ২. রিওয়ার্ড ফাংশন বাদ দেওয়া

DPO-এর মূল ম্যাজিক হলো ওপরের সমীকরণটিকে উল্টে দিয়ে অজানা রিওয়ার্ড $r(x, y)$-কে সরাসরি মডেলের লগ-সম্ভাব্যতা দ্বারা প্রকাশ করা:

r(x, y) = beta log frac{pi^*(y mid x)}{pi_{text{ref}}(y mid x)} + beta log Z(x)

এই মানটিকে ব্র্যাডলি-টেরি সমীকরণে বসিয়ে দিলে জটিল পার্টিশন ফাংশন $Z(x)$ উভয় পাশ থেকে কাটাকাটি হয়ে সম্পূর্ণ বিলুপ্ত হয়ে যায়:

Shakil Mahmud • Research Monograph

P(y_w succ y_l mid x) = sigmaleft( beta log frac{pi^*(y_w mid x)}{pi_{text{ref}}(y_w mid x)} – beta log frac{pi^*(y_l mid x)}{pi_{text{ref}}(y_l mid x)} right)

## ৩. ডিপিও লস ফাংশনের রূপ

এর ফলে আরএলএইচএফের পুরো প্রক্রিয়াটি সাধারণ সুপারভাইজড লার্নিংয়ের মতো একটি সহজ বাইনারি ক্রস-এনট্রপি লস ফাংশনে রূপ নেয়:

mathcal{L}_{text{DPO}}(theta; pi_{text{ref}}) = – mathbb{E}_{(x, y_w, y_l) sim mathcal{D}} left[ log sigmaleft( beta log frac{pi_theta(y_w mid x)}{pi_{text{ref}}(y_w mid x)} – beta log frac{pi_theta(y_l mid x)}{pi_{text{ref}}(y_l mid x)} right) right]

মডেলটি সঠিক বা প্রশংসিত উত্তরের ($y_w$) সম্ভাবনা বাড়িয়ে দেয় এবং বাতিল উত্তরের ($y_l$) সম্ভাবনা কমিয়ে দেয়। কোনো জটিল অ্যাক্টর-ক্রিটিক ট্রেইনিং ছাড়াই DPO ভাষা মডেলকে দ্বিগুণ দ্রুতগতিতে মানুষের পছন্দের সাথে সারিবদ্ধ করতে সক্ষম।

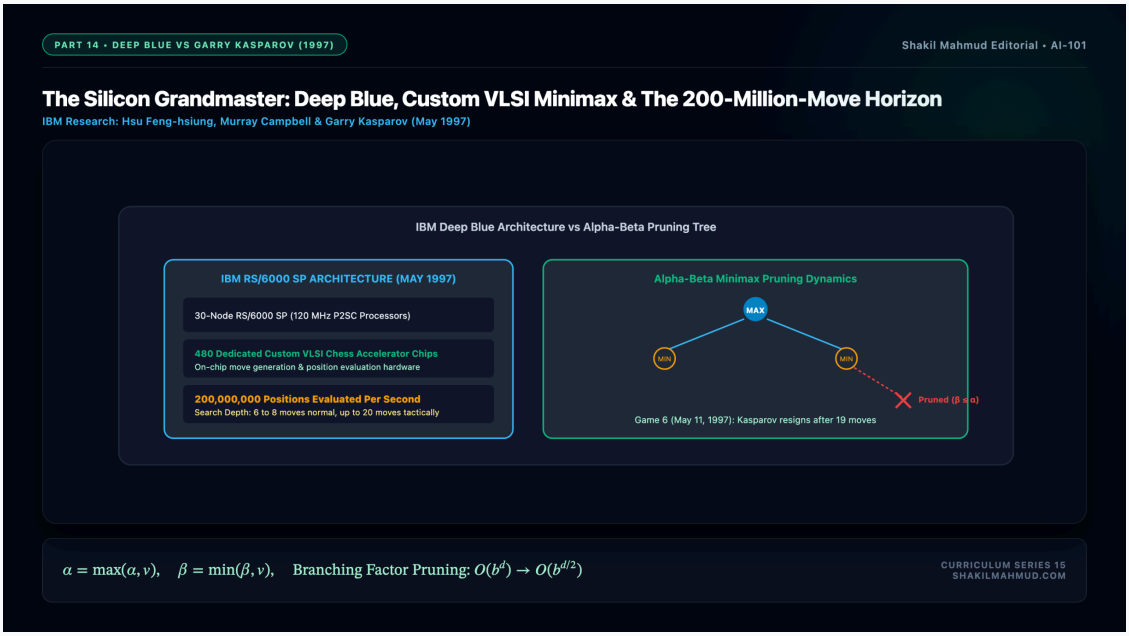
*সাইটেশন: মাহমুদ, শাকিল (2026)। "ডাইরেক্ট প্রেফারেন্স অপটিমাইজেশন (DPO) — রিওয়ার্ড মডেলবিহীন ইমপ্লিসিট লগ-অডস"। লার্জ ল্যাঙ্গুয়েজ মডেল ও জেনারেটিভ এআই, খণ্ড ১৩।*
*https://shakilmahmud.com/bn/large-language-models-part-13-direct-preference-optimization-dpo-implicit-reward-bn/*

শাকিল মাহমুদ মনোগ্রাফ সিরিজ • চতুর্দশ খণ্ড

# প্রম্পট ইঞ্জিনিয়ারিং ও ইন-কনটেক্সট লার্নিং — ইন্ডাকশন হেড ও চেইন-অব-থট যুক্তি

*প্রম্পট ইঞ্জিনিয়ারিং ও ইন-কনটেক্সট লার্নিং — ইন্ডাকশন হেড ও চেইন-অব-থট যুক্তি*

259 শব্দ • November 4, 2026 • ওয়েব সংস্করণ




**চিত্র ১৪:** *আইবিএম ডিপ ব্লুর ৪৮০টি কাস্টম ভিএলএসআই দাবা চিপ এবং কাসপারভের বিরুদ্ধে প্রতি সেকেন্ডে ২০ কোটি অবস্থানের আলফা-বিটা মিনিম্যাক্স সার্চ।*

আধুনিক কৃত্রিম বুদ্ধিমত্তার সবচেয়ে বিস্ময়কর ঘটনাগুলোর একটি হলো **In-Context Learning (ICL)**: মডেলের প্যারামিটারে কোনো পরিবর্তন বা গ্রেডিয়েন্ট আপডেট না ঘটিয়ে কেবল প্রম্পটের ভেতরে কয়েকটি উদাহরণ দেখিয়ে সম্পূর্ণ নতুন কাজ শিখিয়ে ফেলা। অতীতে একে রহস্যময় মনে হলেও মেকানিস্টিক ইন্টারপ্রিটেবিলিটি গবেষণায় উন্মোচিত হয়েছে এর পেছনের নিউরাল সার্কিট—বিশেষ করে **ইন্ডাকশন হেডস (Induction Heads)** এবং লুকায়িত মেটা-অপটিমাইজেশন গতিবিদ্যা।

## ১. ইন্ডাকশন হেডসের অ্যানাটমি

২০২২ সালে অ্যানথ্রপিকের গবেষকরা আবিষ্কার করেন যে ট্রান্সফরমারে দুই স্তরবিশিষ্ট কিছু বিশেষ অ্যাটেনশন হেড থাকে যা প্যাটার্ন কপি করার অ্যালগরিদম চালায়:

$$dots[A][B]dots[A]longrightarrow[B]$$

প্রম্পটের কোথাও যদি টোকেন $[A]$-এর পর $[B]$ থাকে, এবং পরে আবার $[A]$ ফিরে আসে, তবে এই হেডটি সরাসরি অতীত থেকে $[B]$-এর মান এনে পরবর্তী পূর্বাভাস হিসেবে প্রদান করে। এই দুটি স্তরের সমন্বিত কাজ কোটি কোটি শব্দ ও ব্যাকরণগত ধারণার ওপর একযোগে কার্যকর হয়ে ইন-কনটেক্সট শেখার ভিত্তি তৈরি করে।

## ২. অন্তর্নিহিত গ্রেডিয়েন্ট ডিসেন্ট

গবেষণায় দেখা গেছে যে ট্রান্সফরমারের সেলফ-অ্যাটেনশন ম্যাট্রিক্স ইনফারেন্সের সময় প্রম্পটের উদাহরণগুলোকে ডেটা হিসেবে নিয়ে অভ্যন্তরীণভাবে **গ্রেডিয়েন্ট ডিসেন্টের মতো কাজ করে**। ফলে প্রম্পটে দেওয়া নির্দেশনাগুলো মডেলের ভেতর একধরনের ক্ষণস্থায়ী নতুন ওজন সমন্বয় তৈরি করে।

## ৩. চেইন-অব-থট (Chain-of-Thought) ও যুক্তিশৃঙ্খল

একটি ট্রান্সফরমার প্রতি টোকেনের জন্য একটি নির্দিষ্ট পরিমাণ কম্পিউটেশন বা ফ্লপস খরচ করতে পারে। কোনো জটিল গাণিতিক সমস্যার সরাসরি এক শব্দে উত্তর দিতে বললে মডেলের পক্ষে একটিমাত্র ফরোয়ার্ড পাসে পুরো হিসাব সমাধান করা অসম্ভব হয়ে পড়ে।

Shakil Mahmud • Research Monograph

**“ধাপে ধাপে চিন্তা করো” (Chain-of-Thought)** কৌশল ব্যবহার করলে মডেলটি উত্তর দেওয়ার আগে মধ্যবর্তী যুক্তির টোকেন তৈরি করে। প্রতিটি টোকেন তার KV ক্যাশে যুক্ত হয় এবং পরবর্তী পদক্ষেপের জন্য একটি সহায়ক মেমরি বা ভ্যারিয়েবল হিসেবে কাজ করে, যার ফলে মডেলের জটিল যুক্তি ও গণিতের নির্ভুলতা বহুগুণ বৃদ্ধি পায়।



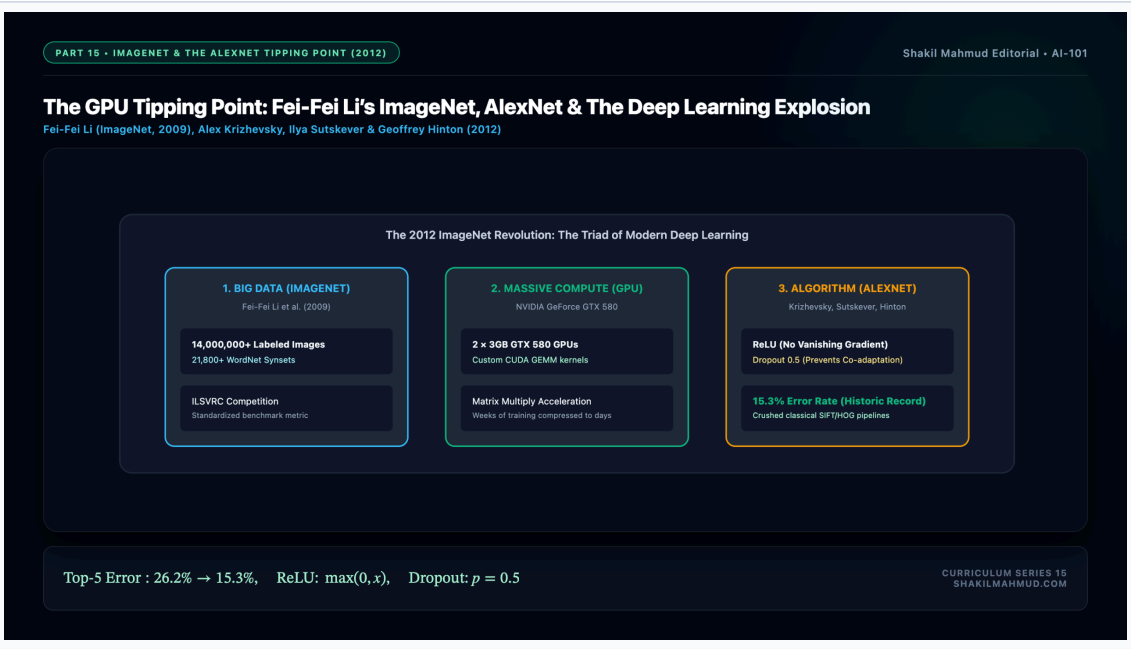
*সাইটেশন: মাহমুদ, শাকিল (2026)। "প্রম্পট ইঞ্জিনিয়ারিং ও ইন-কনটেক্সট লার্নিং — ইন্ডাকশন হেড ও চেইন-অব-থট যুক্তি"। লার্জ ল্যাঙ্গুয়েজ মডেল ও জেনারেটিভ এআই, খণ্ড ১৪।*
*https://shakilmahmud.com/bn/large-language-models-part-14-prompt-engineering-in-context-learning-induction-heads-bn/*

শাকিল মাহমুদ মনোগ্রাফ সিরিজ • পঞ্চদশ খণ্ড

# রিট্রিভাল-অগমেন্টেড জেনারেশন (RAG) — ডেন্স এম্বেডিং, HNSW ইনডেক্সিং ও হাইব্রিড রির‍্যাংকিং

*রিট্রিভাল-অগমেন্টেড জেনারেশন (RAG) — ডেন্স এম্বেডিং, HNSW ইনডেক্সিং ও হাইব্রিড রির‍্যাংকিং*

271 শব্দ • November 4, 2026 • ওয়েব সংস্করণ




**চিত্র ১৫:** *২০১২ সালের যুগান্তকারী ডিপ লার্নিং বিস্ফোরণ: ফেই-ফেই লির ইমেজনেট ডেটাসেট, জিপিইউ প্যারালালেলিজম এবং অ্যালেক্সনেটের রিলু ও ড্রপআউট।*

বিশাল মেধা থাকা সত্ত্বেও লার্জ ল্যাঙ্গুয়েজ মডেলগুলোর মেমরি নির্দিষ্ট সময়ে সীমাবদ্ধ থাকে এবং তাদের সম্ভাব্যতামূলক স্বভাবের কারণে তারা প্রায়শই **হ্যালুসিনেশন (Hallucination)** বা অলীক তথ্য তৈরি করে। **রিট্রিভাল-অগমেন্টেড জেনারেশন (RAG)** ভাষা মডেলের মেধার সাথে বহিরাগত বিশ্বস্ত নথিপত্র যুক্ত করে দেয়। ব্যবহারকারীর প্রশ্নের সাথে সম্পর্কিত বাস্তব তথ্য কোটি কোটি দলিলের মধ্য থেকে তাৎক্ষণিকভাবে খুঁজে এনে মডেলের সামনে উপস্থাপন করা হয়, যার ফলে মডেলের প্রতিটি উত্তর প্রমাণভিত্তিক ও নির্ভরযোগ্য হয়।

## ১. RAG আর্কিটেকচারের তিনটি মূল স্তম্ভ

২০২০ সালে প্যাট্রিক লুইস প্রমুখ RAG আর্কিটেকচার প্রস্তাব করেন:

1. **নথি রূপান্তর ও ভেক্টরাইজেশন:** যেকোনো দীর্ঘ টেক্সটকে ক্ষুদ্র খণ্ডে ভাগ করে একটি এম্বেডিং মডেল দিয়ে গাণিতিক ভেক্টরে রূপান্তর করা হয়।
2. **দ্রুত অনুসন্ধান (ANN Retrieval):** ব্যবহারকারীর প্রশ্নকেও ভেক্টরে রূপান্তর করে ভেক্টর ডেটাবেস থেকে সবচেয়ে কাছাকাছি অর্থবোধক নথিগুলো খুঁজে বের করা হয়।
3. **প্রমাণভিত্তিক উত্তর প্রদান:** উদ্ধারকৃত বাস্তব তথ্যগুলোকে প্রম্পটের সাথে যুক্ত করে মডেলকে উত্তর তৈরি করার নির্দেশ দেওয়া হয়।

## ২. HNSW গ্রাফ ইনডেক্সিংয়ের বিস্ময়

লাখ লাখ ভেক্টরের সাথে প্রতিটি তুলনামূলক দূরত্ব মাপা অত্যন্ত সময়সাপেক্ষ ($O(N)$)। **HNSW (Hierarchical Navigable Small World)** গ্রাফ অ্যালগরিদম স্কিপ-লিস্টের মতো একাধিক স্তরের গ্রাফ তৈরি করে। ওপরের স্তরে থাকে দীর্ঘ দূরত্বের সংযোগ, যার ফলে অনুসন্ধান অত্যন্ত দ্রুত সঠিক ক্লাস্টারের কাছে পৌঁছে যায়। এরপর নিচের স্তরে নেমে নিখুঁত প্রতিবেশী খুঁজে বের করে। ফলে $O(logN)$ সময়ে ৯৮%-এর বেশি নিখুঁত তথ্য উদ্ধার করা সম্ভব হয়।

Shakil Mahmud • Research Monograph

## ৩. হাইব্রিড সার্চ ও ক্রস-এনকোডার রির‍্যাংকিং

অর্থবোধক ভেক্টর সার্চ অনেক সময় সুনির্দিষ্ট কোড নম্বর বা অদ্ভুত নামের ক্ষেত্রে ভুল করতে পারে। আধুনিক RAG পাইপলাইনগুলো **হাইব্রিড সার্চ** ব্যবহার করে—যেখানে ভেক্টর সার্চ এবং প্রাচীন কিওয়ার্ড সার্চ (BM25) উভয়কে একত্রিত করা হয়। উদ্ধারকৃত শীর্ষ ১০০টি নথিকে একটি গভীর **Cross-Encoder** দিয়ে পুনরায় মূল্যায়ন ও সাজানো হয়, যার ফলে সবচেয়ে প্রাসঙ্গিক ও নিখুঁত তথ্যটিই কেবল ভাষা মডেলের ইনপুটে পৌঁছায়।



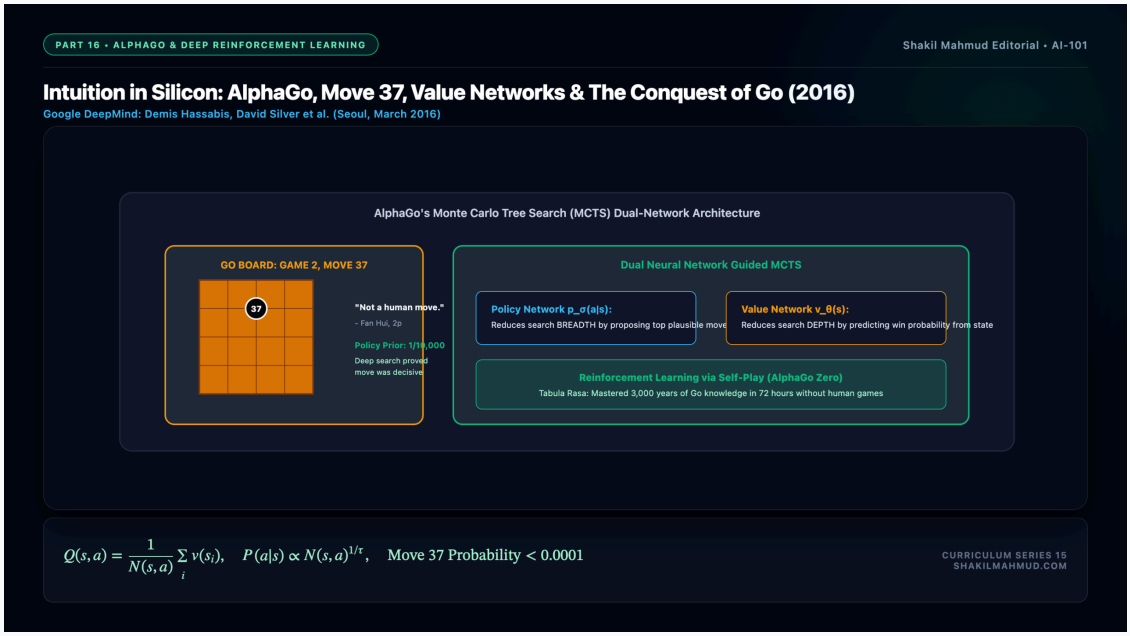
*সাইটেশন: মাহমুদ, শাকিল (2026)। "রিট্রিভাল-অগমেন্টেড জেনারেশন (RAG) — ডেন্স এম্বেডিং, HNSW ইনডেক্সিং ও হাইব্রিড রির‍্যাংকিং"। লার্জ ল্যাঙ্গুয়েজ মডেল ও জেনারেটিভ এআই, খণ্ড ১৫। https://shakilmahmud.com/bn/large-language-models-part-15-retrieval-augmented-generation-rag-dense-embeddings-hnsw-bn/*

শাকিল মাহমুদ মনোগ্রাফ সিরিজ • ষোড়শ খণ্ড

# অটোরিগ্রেসিভ ডিকোডিং কৌশল — টেম্পারেচার, টপ-পি নিউক্লিয়াস ও স্পেকুলেটিভ ডিকোডিং

*অটোরিগ্রেসিভ ডিকোডিং কৌশল — টেম্পারেচার, টপ-পি নিউক্লিয়াস ও স্পেকুলেটিভ ডিকোডিং*

311 শব্দ • November 4, 2026 • ওয়েব সংস্করণ




**চিত্র ১৬:** *আলফাগোর পলিসি ও ভ্যালু নেটওয়ার্ক চালিত মন্টে কার্লো ট্রি সার্চ এবং ঐতিহাসিক চাল ৩৭ (Move 37) এর কম্পিউটার অন্তর্দৃষ্টি।*

একটি লার্জ ল্যাঙ্গুয়েজ মডেলের শেষ স্তর সরাসরি কোনো শব্দ দেয় না; এটি পুরো শব্দভাণ্ডারের জন্য আননরমালাইজড লগিট ভেক্টর $mathbf{z} in mathbb{R}^{|mathcal{V}|}$ তৈরি করে। এই অবিচ্ছিন্ন সংখ্যাগুলোকে কীভাবে বাস্তব বাক্যের শব্দে রূপান্তর করা হবে—সেই **ডিকোডিং কৌশল (Decoding Strategy)** নির্ধারণ করে মডেলটি কি একই কথা বারবার পুনরাবৃত্তি করবে, নাকি সৃজনশীল ও সুসংগত আলোচনা উপহার দেবে। তাছাড়া আধুনিক **স্পেকুলেটিভ ডিকোডিং** কোনো গুণগত সমঝোতা ছাড়াই টেক্সট তৈরির গতি দ্বিগুণ থেকে তিনগুণ বাড়িয়ে দেয়।

## ১. তাপমাত্রার থার্মোডাইনামিক্স (Temperature Scaling)

সফটম্যাক্স সূত্রে টেম্পারেচার $T$ যোগ করে সম্ভাব্যতা বণ্টনকে নিয়ন্ত্রণ করা হয়:

P(w_i) = frac{exp(z_i / T)}{sum_{j=1}^{|mathcal{V}|} exp(z_j / T)}

- $T to 0$ **(Greedy Search):** সর্বদা সর্বোচ্চ স্কোরের শব্দটিই বেছে নেওয়া হয়। গণিত ও কোডিংয়ের জন্য এটি সেরা হলেও দীর্ঘ সংলাপে এটি বৃত্তাকার পুনরাবৃত্তি ঘটায়।
- $T = 1.0$: মডেলের প্রাকৃতিক সম্ভাব্যতা বণ্টন।
- **T > 1.0**: বণ্টনকে সমান ও চ্যাপ্টা করে তোলে, যার ফলে বিচিত্র ও সৃজনশীল শব্দ আসে, তবে অতিরিক্ত হলে বাক্য অর্থহীন হয়ে পড়ে।

## ২. টপ-$k$ বনাম টপ-$p$ (নিউক্লিয়াস) স্যাম্পলিং

শুধু তাপমাত্রা বাড়ালে শব্দভাণ্ডারের পেছনের অদ্ভুত অপ্রাসঙ্গিক শব্দ চলে আসার ঝুঁকি থাকে।

- **টপ-$k$:** কেবল শীর্ষ $k$-টি শব্দের মধ্যে থেকে বাছাই করে। কিন্তু ভাষা সব সময় সমান অনিশ্চিত থাকে না।

Shakil Mahmud • Research Monograph

- **টপ-$p$ (নিউক্লিয়াস স্যাম্পলিং):** সঞ্চিত সম্ভাব্যতা যোগ করে যতক্ষণ পর্যন্ত $p$ (যেমন ০.৯) পূর্ণ না হয়, ঠিক ততক্ষণ পর্যন্ত সেরা শব্দগুলোকে তালিকায় রাখে। উত্তর যখন নিশ্চিত তখন তালিকা ছোট হয়ে যায়, আর যখন বিকল্প অনেক তখন তালিকা বড় হয়।

## ৩. স্পেকুলেটিভ ডিকোডিং (Speculative Decoding)

চিরায়ত নিয়মে ৭০ বিলিয়ন মডেল দিয়ে একটি একটি করে টোকেন তৈরি করা খুবই ধীরগতির। স্পেকুলেটিভ পদ্ধতিতে একটি ক্ষুদ্র ও ক্ষিপ্র ১ বিলিয়ন মডেল দ্রুত ৫টি টোকেনের খসড়া তৈরি করে ফেলে। এরপর মূল ৭০ বিলিয়ন মডেলটি একটিমাত্র ফরোয়ার্ড পাসে ৫টি টোকেন একসাথে যাচাই করে নেয়:

P(text{accept } hat{x}_i) = minleft(1, frac{P_{text{target}}(hat{x}_i)}{P_{text{draft}}(hat{x}_i)}right)

যেহেতু বড় মডেলের জন্য ১টি টোকেন তৈরি করা এবং ৫টি টোকেন একসাথে ভেরিফাই করার সময় প্রায় সমান, তাই এই প্রক্রিয়ায় গুণমানের চুল পরিমাণ ক্ষতি না করেই ইনফারেন্সের গতি ২ থেকে ৩ গুণ বৃদ্ধি পায়।



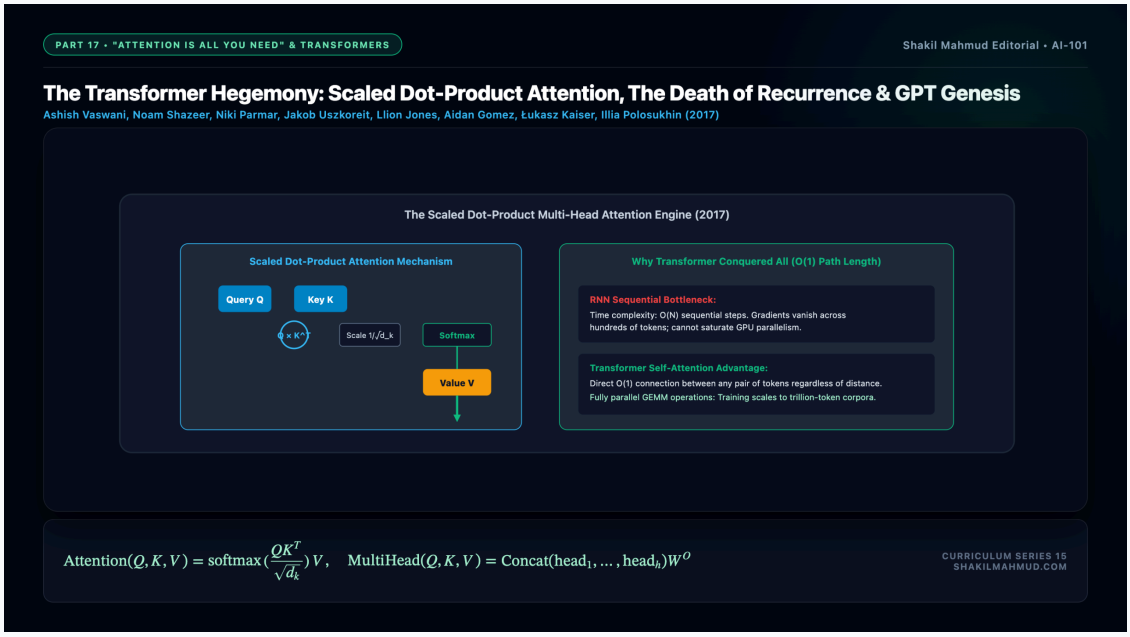
*সাইটেশন: মাহমুদ, শাকিল (2026)। "অটোরিগ্রেসিভ ডিকোডিং কৌশল — টেম্পারেচার, টপ-পি নিউক্লিয়াস ও স্পেকুলেটিভ ডিকোডিং"। লার্জ ল্যাঙ্গুয়েজ মডেল ও জেনারেটিভ এআই, খণ্ড ১৬।*
*https://shakilmahmud.com/bn/large-language-models-part-16-autoregressive-decoding-temperature-top-p-speculative-bn/*

শাকিল মাহমুদ মনোগ্রাফ সিরিজ • সপ্তদশ খণ্ড

# স্কোর-ভিত্তিক কন্টিনিউয়াস ডিফিউশন — স্টোকাস্টিক ডিফারেনশিয়াল সমীকরণ ও DDPM

*স্কোর-ভিত্তিক কন্টিনিউয়াস ডিফিউশন — স্টোকাস্টিক ডিফারেনশিয়াল সমীকরণ ও DDPM*

280 শব্দ • November 4, 2026 • ওয়েব সংস্করণ




**চিত্র ১৭:** *ট্রান্সফরমার আর্কিটেকচার: স্কেলড ডট-প্রোডাক্ট মাল্টি-হেড সেলফ-অ্যাটেনশন, যা রিকারেন্ট বাধা দূর করে O(১) টোকেন সংযোগ স্থাপন করেছে।*

অটোরিগ্রেসিভ ট্রান্সফরমার ডিসক্রিট টেক্সট তৈরিতে রাজত্ব করলেও, বাস্তবসম্মত ছবি, ভিডিও ও অডিও তৈরির জগতে সবচেয়ে শক্তিশালী আর্কিটেকচার হলো **ডিফিউশন মডেল (Diffusion Models)**। নন-ইকুইলিব্রিয়াম থার্মোডাইনামিক্সের ওপর ভিত্তি করে গড়ে ওঠা এই পদ্ধতিতে সৃষ্টিকে দেখা হয় পদার্থবিজ্ঞানের এনট্রপির বিপরীত প্রক্রিয়া হিসেবে: একটি পরিষ্কার ছবিতে ধাপে ধাপে গাউসিয়ান নয়েজ যোগ করে পুরোপুরি বিশৃঙ্খলায় রূপান্তর করা হয় এবং একটি নিউরাল নেটওয়ার্ককে শেখানো হয় কীভাবে সময়ের কাঁটাকে উল্টো ঘুরিয়ে সেই বিশৃঙ্খলা থেকে নিখুঁত ছবি পুনর্নির্মাণ করতে হয়।

## ১. ফরোয়ার্ড নয়েজিং প্রক্রিয়া (DDPM)

২০২০ সালে জনাথন হো প্রমুখ DDPM মডেল প্রস্তাব করেন। একটি মূল ছবি $mathbf{x}_0$-তে $T$ সংখ্যক ধাপে সামান্য করে নয়েজ যোগ করা হয়:

q(mathbf{x}_t mid mathbf{x}_{t-1}) = mathcal{N}left(mathbf{x}_t; sqrt{1 – beta_t} mathbf{x}_{t-1}, beta_t mathbf{I}right)

রিপ্যারামিটারাইজেশন ট্রিকের কল্যাণে যেকোনো ধাপ $t$-এর নয়েজযুক্ত রূপ সরাসরি একটিমাত্র সমীকরণে বের করা যায়:

mathbf{x}_t = sqrt{bar{alpha}_t} mathbf{x}_0 + sqrt{1 – bar{alpha}_t} boldsymbol{epsilon}, quad boldsymbol{epsilon} sim mathcal{N}(mathbf{0}, mathbf{I})

শেষ ধাপে পৌঁছালে ছবিটি সম্পূর্ণরূপে অর্থহীন গাউসিয়ান নয়েজে পরিণত হয়।

## ২. রিভার্স প্রসেস ও স্কোর ম্যাচিং

ছবি তৈরির জন্য এলোমেলো নয়েজ থেকে শুরু করে উল্টো পথে ফিরে আসতে হয়। ইয়াং সং এবং স্টেফানো এরমন দেখান যে এই উল্টো যাত্রা পরিচালনা করার জন্য ডেটা স্পেসের **স্টাইন স্কোর ফাংশন (Stein Score Function)** $nabla_{mathbf{x}} log q_t(mathbf{x})$ শেখা প্রয়োজন—যা নির্দেশ করে কোন দিকে গেলে নয়েজ কমে ছবির স্পষ্টতা বাড়ে।

Shakil Mahmud • Research Monograph

## ৩. কন্টিনিউয়াস স্টোকাস্টিক ডিফারেনশিয়াল সমীকরণ (SDE)

সময়কে অবিচ্ছিন্ন ধারা হিসেবে ধরে ডিফিউশনকে ইতো স্টোকাস্টিক ডিফারেনশিয়াল সমীকরণ (SDE) দিয়ে প্রকাশ করা হয়:

mathrm{d}mathbf{x} = mathbf{f}(mathbf{x}, t) mathrm{d}t + g(t) mathrm{d}mathbf{w}

একটি নিউরাল নেটওয়ার্ক $boldsymbolepsilon_theta(mathbf x_t, t)$ প্রশিক্ষণ দেওয়া হয় ছবির ভেতরে যোগ করা নয়েজ অনুমান করার জন্য:

mathcal{L}(theta) = mathbb{E} left[ | boldsymbol{epsilon} – boldsymbol{epsilon}_theta(mathbf{x}_t, t) |_2^2 right]

প্রশিক্ষণ শেষে ওডিই সলভারের (ODE Solvers) মাধ্যমে এলোমেলো নয়েজের জটলা থেকে অবিশ্বাস্য স্পষ্টতার ছবি ও শিল্পকর্ম সৃষ্টি করা সম্ভব হয়।

---

*সাইটেশন: মাহমুদ, শাকিল (2026)। "স্কোর-ভিত্তিক কন্টিনিউয়াস ডিফিউশন — স্টোকাস্টিক ডিফারেনশিয়াল সমীকরণ ও DDPM"। লার্জ ল্যাঙ্গুয়েজ মডেল ও জেনারেটিভ এআই, খণ্ড ১৭।*
*https://shakilmahmud.com/bn/large-language-models-part-17-score-based-continuous-generative-diffusion-sde-ddpm-bn/*

শাকিল মাহমুদ মনোগ্রাফ সিরিজ • অষ্টাদশ খণ্ড

# ল্যাটেন্ট ডিফিউশন ও ফ্লো ম্যাচিং — স্টেবল ডিফিউশন ও রেকটিফায়েড ফ্লোস

*ল্যাটেন্ট ডিফিউশন ও ফ্লো ম্যাচিং — স্টেবল ডিফিউশন ও রেকটিফায়েড ফ্লোস*

265 শব্দ • November 4, 2026 • ওয়েব সংস্করণ




**চিত্র ১৮:** *নিউরাল স্কেলিং সূত্র: কাপলান ও চিনচিলা পাওয়ার-ল সীমান্ত এবং জিপিটি-৩ তে ইন-কনটেক্সট ফিউ-শট প্রম্পটিংয়ের আত্মপ্রকাশ।*

সরাসরি উচ্চমানের পিক্সেলের ওপর (১০২৪$times$১০২৪$times$৩$approx$৩০ লাখ ডাইমেনশন) ডিফিউশন পরিচালনা করা অসম্ভব রকমের ধীর ও ব্যয়বহুল। **ল্যাটেন্ট ডিফিউশন মডেল (Latent Diffusion Models / Stable Diffusion)** এবং আধুনিক **ফ্লো ম্যাচিং (Flow Matching)** পুরো প্রক্রিয়াকে দুটি ভাগে বিভক্ত করে: বাহ্যিক অনুভূতির সংকোচন এবং অভ্যন্তরীণ অর্থবোধক সৃষ্টি—যার ফলে ব্যক্তিগত কম্পিউটারেও চোখের পলকে অভূতপূর্ব শিল্পকর্ম তৈরি করা সম্ভব হয়েছে।

## ১. ল্যাটেন্ট স্পেসের দ্বি-স্তরীয় বিভাজন

২০২২ সালে রমবার্ক প্রমুখ আবিষ্কার করেন যে ছবি তৈরিতে দুটি সম্পূর্ণ ভিন্ন জগত থাকে:

1. **অনুভূতিগত সংকোচন (Perceptual Compression):** একটি VAE অটোএনকোডার দিয়ে বিশাল পিক্সেল ছবিকে ৬৪ গুণ সংকুচিত করে একটি অতি ক্ষুদ্র ল্যাটেন্ট স্পেসে $mathbf{z}$ রূপান্তর করা হয়। এতে চোখে দেখা যায় না এমন অপ্রয়োজনীয় সূক্ষ্ম নয়েজ দূর হয়ে যায়।
2. **অর্থবোধক সৃষ্টি (Semantic Synthesis):** ডিফিউশন মডেলটি (U-Net বা DiT) কেবল এই ক্ষুদ্র ল্যাটেন্ট স্পেসের ভেতরে কাজ করে, ফলে কম্পিউটেশনাল খরচ কয়েকগুণ কমে যায়। শেষে VAE ডিকোডার দিয়ে পূর্ণাঙ্গ ছবি ফিরিয়ে আনা হয়।

## ২. ক্লাসিফায়ার-ফ্রি গাইডেন্স (CFG)

ব্যবহারকারীর টেক্সট প্রম্পটের সাথে ছবির মিল নিশ্চিত করতে **Classifier-Free Guidance (CFG)** ব্যবহার করা হয়। টেক্সটযুক্ত পূর্বাভাস এবং টেক্সটবিহীন খালি পূর্বাভাসের মধ্যকার পার্থক্যকে স্কেল $s$ দিয়ে বাড়িয়ে দেওয়া হয়:

tilde{boldsymbol{epsilon}} = boldsymbol{epsilon}(emptyset) + s cdot (boldsymbol{epsilon}(mathbf{c}) – boldsymbol{epsilon}(emptyset))

এর ফলে ছবির বিষয়বস্তু ব্যবহারকারীর প্রম্পটের প্রতিটি শব্দের সাথে কঠোরভাবে মিলে যায়।

Shakil Mahmud • Research Monograph

## ৩. ফ্লো ম্যাচিং ও রেকটিফায়েড সোজা গতিপথ

চিরায়ত ডিফিউশনে এলোমেলো বাঁকা পথে নয়েজ সরানো হতো, ফলে ৩০ থেকে ৫০টি পদক্ষেপ লাগত। **Flow Matching** এবং **Rectified Flows** নয়েজ থেকে ছবির পরিবর্তনকে একদম সরলরেখায় সম্পন্ন করে:

mathbf{x}_t = (1 – t)mathbf{x}_0 + t mathbf{x}_1

গতিপথ সরলরেখা হওয়ায় মাত্র **১ থেকে ৪টি ধাপে** সম্পূর্ণ নিখুঁত ছবি তৈরি করা সম্ভব হয়—যা Stable Diffusion 3 এবং Flux-এর মূল প্রযুক্তি।

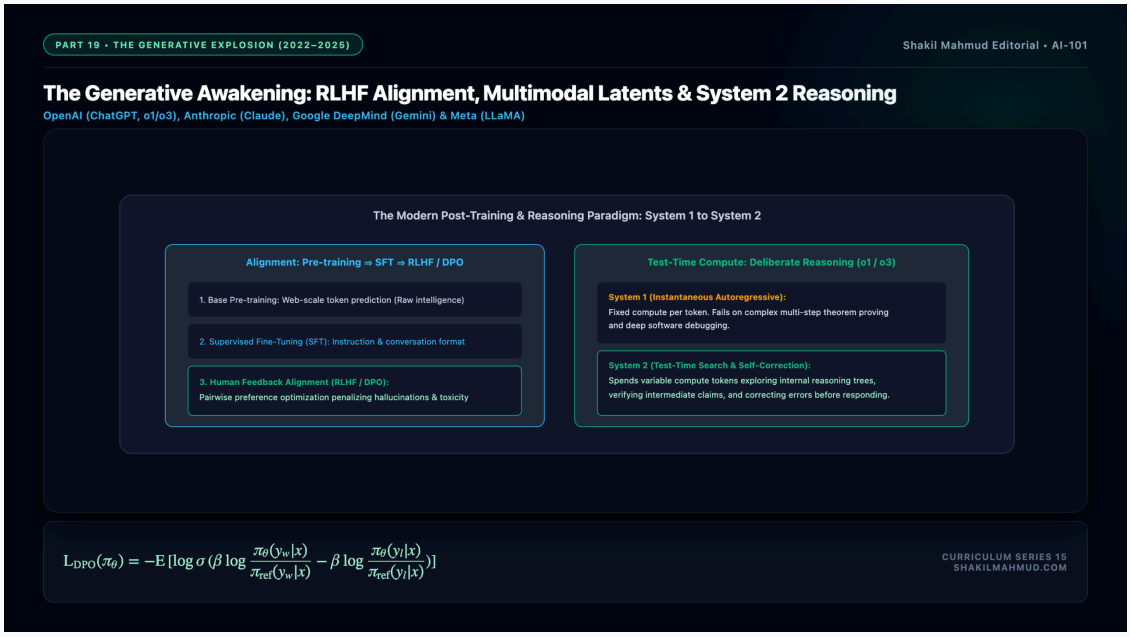
*সাইটেশন: মাহমুদ, শাকিল (2026)। "ল্যাটেন্ট ডিফিউশন ও ফ্লো ম্যাচিং — স্টেবল ডিফিউশন ও রেকটিফায়েড ফ্লোস"। লার্জ ল্যাঙ্গুয়েজ মডেল ও জেনারেটিভ এআই, খণ্ড ১৮।*
*https://shakilmahmud.com/bn/large-language-models-part-18-latent-diffusion-flow-matching-stable-diffusion-rectified-bn/*

শাকিল মাহমুদ মনোগ্রাফ সিরিজ • ঊনবিংশ খণ্ড

# লং-কনটেক্সট সিস্টেম — রিং-অ্যাটেনশন, RoPE ফ্রিকোয়েন্সি স্কেলিং ও নিডল-ইন-আ-হেস্ট্যাক

*লং-কনটেক্সট সিস্টেম — রিং-অ্যাটেনশন, RoPE ফ্রিকোয়েন্সি স্কেলিং ও নিডল-ইন-আ-হেস্ট্যাক*

304 শব্দ • November 4, 2026 • ওয়েব সংস্করণ




**চিত্র ১৯:** *আধুনিক জেনারেটিভ এআই বিপ্লব: আরএলএইচএফ ও ডিপিও অ্যালাইনমেন্ট থেকে সিস্টেম-২ টেস্ট-টাইম রিজনিং সার্চের রূপরেখা।*

আদিম ল্যাঙ্গুয়েজ মডেলগুলোর কনটেক্সট উইন্ডো মাত্র ২,০৪৮ টোকেনে সীমাবদ্ধ ছিল—যা একটি ছোট প্রবন্ধের সমান। আজ অত্যাধুনিক এআই মডেলগুলো অনায়াসে এক বা দুই মিলিয়ন টোকেন একবারে ধারণ করতে পারে, যা পুরো লাইব্রেরি, শত শত কোড ফাইল বা ঘণ্টার পর ঘণ্টা ভিডিওর সমতুল্য। এই সুবিশাল দৈর্ঘ্যে পৌঁছাতে সেলফ-অ্যাটেনশনের মেমরি বিস্ফোরণ এবং দূরত্বের ভুল প্রতিরোধে উদ্ভাবিত হয়েছে **রিং-অ্যাটেনশন (RingAttention)** এবং **RoPE ফ্রিকোয়েন্সি স্কেলিং**।

## ১. RoPE ফ্রিকোয়েন্সি স্কেলিং

রোটারি পজিশনাল এম্বেডিংয়ে (RoPE) শব্দের অবস্থান কোণের আবর্তনের মাধ্যমে বোঝানো হয়। ৪,০০০ টোকেনের জন্য তৈরি মডেলকে সরাসরি ৩২,০০০ টোকেনে চালালে নতুন কোণের মান দেখে মডেলের সমস্ত হিসাব ভেঙে পড়ে।

নতুন করে পুরো প্রশিক্ষণ না দিয়ে গণিতের মাধ্যমে একে বিস্তৃত করা হয়:

- **NTK-Aware স্কেলিং:** মূল বেস ফ্রিকোয়েন্সি $b$-কে সমন্বয় করে উচ্চ-ফ্রিকোয়েন্সির স্থানীয় ব্যাকরণকে ঠিক রাখা হয় এবং নিম্ন-ফ্রিকোয়েন্সির দীর্ঘ দূরত্বের কোণকে সংকুচিত করা হয়।
- **YaRN পদ্ধতি:** তাপমাত্রা সমন্বয়ের মাধ্যমে বহু মিলিয়ন টোকেন দৈর্ঘ্যেও মনোযোগ নষ্ট হতে দেয় না।

## ২. রিং-অ্যাটেনশন (RingAttention): বৃত্তাকার মেমরি বিতরণ

১০ লাখ টোকেনের Key-Value ক্যাশ একটিমাত্র জিপিইউর মেমরিতে রাখা অসম্ভব (১০০ জিবির বেশি)। ২০২৩ সালে ইউসি বার্কলের গবেষকরা উদ্ভাবন করেন **RingAttention**। এখানে একাধিক জিপিইউকে একটি বৃত্তাকার বা রিং নেটওয়ার্কে যুক্ত করা হয়:

1. প্রতিটি জিপিইউ সম্পূর্ণ লেখার একটি অংশ ধারণ করে।

Shakil Mahmud • Research Monograph

2. নিজের অংশের হিসাব করার পাশাপাশি প্রতিটি জিপিইউ ব্যাকগ্রাউন্ডে তার ডেটা পাশের জিপিইউকে পাঠিয়ে দেয় এবং অপর পাশের জিপিইউ থেকে নতুন ডেটা গ্রহণ করে।
3. অনলাইন সফটম্যাক্সের সাহায্যে ধাপে ধাপে পুরো বৃত্তের তথ্য সমন্বয় করা হয়।

যোগাযোগ ও গণনাকর্ম একসাথে চলায় মেমরির সীমাবদ্ধতা দূর হয়ে যায় এবং ক্লাস্টারে জিপিইউ যোগ করে কনটেক্সট উইন্ডোকে কোটি টোকেন পর্যন্ত বিস্তৃত করা সম্ভব হয়।

## ৩. নিডল-ইন-আ-হেস্ট্যাক (Needle In A Haystack) পরীক্ষা

মডেল সত্যি পুরো লেখাটি পড়ছে নাকি মাঝের অংশ বাদ দিয়ে দিচ্ছে তা যাচাই করতে ১০ লাখ শব্দের বিশাল উপন্যাসের কোনো এক অজানা কোণায় একটি গোপন বাক্য লুকিয়ে রেখে প্রশ্ন করা হয়। আধুনিক ফ্রন্টিয়ার মডেলগুলো শূন্য থেকে শতভাগ গভীরতার যেকোনো স্থান থেকে শতভাগ নির্ভুলভাবে তথ্য উদ্ধার করে তাদের কার্যকর স্মৃতিশক্তির প্রমাণ দিয়েছে।

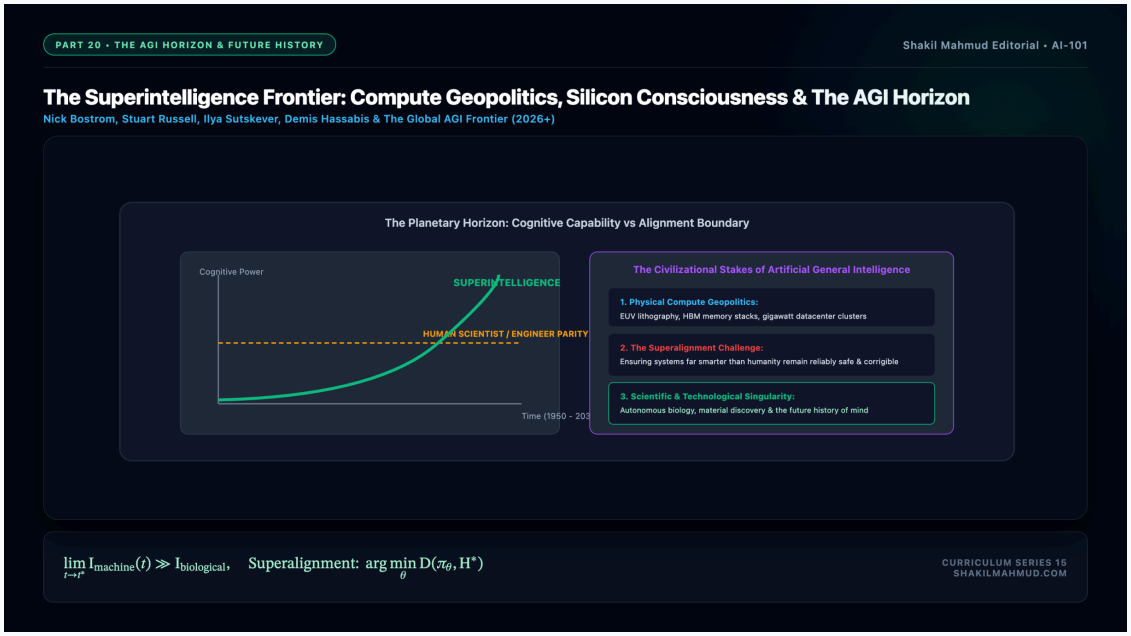
*সাইটেশন: মাহমুদ, শাকিল (2026)। "লং-কনটেক্সট সিস্টেম — রিং-অ্যাটেনশন, RoPE ফ্রিকোয়েন্সি স্কেলিং ও নিডল-ইন-আ-হেস্ট্যাক"। লার্জ ল্যাঙ্গুয়েজ মডেল ও জেনারেটিভ এআই, খণ্ড ১৯।*
*https://shakilmahmud.com/bn/large-language-models-part-19-long-context-systems-ringattention-rope-needle-haystack-bn/*

শাকিল মাহমুদ মনোগ্রাফ সিরিজ • বিংশ খণ্ড

# কৃত্রিম প্রজ্ঞার দিগন্ত — মাল্টিমডালিটি, রিজনিং টোকেন ও টেস্ট-টাইম কম্পিউট

*কৃত্রিম প্রজ্ঞার দিগন্ত — মাল্টিমডালিটি, রিজনিং টোকেন ও টেস্ট-টাইম কম্পিউট*

280 শব্দ • November 4, 2026 • ওয়েব সংস্করণ




**চিত্র ২০:** *কৃত্রিম সাধারণ বুদ্ধিমত্তা (AGI) দিগন্ত: সূচকীয় সক্ষমতা বৃদ্ধি, সেমিকন্ডাক্টর কম্পিউট অবকাঠামো এবং সুপারঅ্যালাইনমেন্টের চ্যালেঞ্জ।*

জেনারেটিভ এআইয়ের ঐতিহাসিক অগ্রযাত্রা আজ এক অভূতপূর্ব মোহনায় এসে উপনীত হয়েছে: টেক্সট, দৃষ্টি, শব্দ ও গভীর যুক্তিশৃঙ্খলের মধ্যকার সমস্ত কৃত্রিম সীমানা মুছে গেছে। আধুনিক ফ্রন্টিয়ার মডেলগুলো কেবল ভাষার অনুকরণকারী নয়; বরং তারা পরিণত হয়েছে সমন্বিত মাল্টিমোডাল প্রজ্ঞার ইঞ্জিনে। ইনফারেন্সের সময় অতিরিক্ত কম্পিউটেশন ব্যয় করে **রিজনিং টোকেন (Reasoning Tokens)** ও **টেস্ট-টাইম সার্চের** মাধ্যমে তারা প্রি-ট্রেনিংয়ের সীমাবদ্ধতা ভেঙে নতুন দিগন্ত উন্মোচন করেছে।

## ১. একীভূত মাল্টিমোডাল এম্বেডিং জগত

আধুনিক মডেলগুলো (যেমন GPT-4o বা Gemini) পৃথক কোনো মডিউল ছাড়াই সরাসরি ছবি ও শব্দকে টোকেন হিসেবে গ্রহণ করে। একটি বিশেষ প্রজেক্টরের মাধ্যমে ছবির ভিজ্যুয়াল প্যাচগুলোকে সরাসরি ভাষা মডেলের ভেক্টর স্পেসে ম্যাপ করা হয়। ফলে সেলফ-অ্যাটেনশন লেয়ারগুলো একই সাথে ছবি, কোড এবং টেক্সটের মধ্যে নিবিড় সম্পর্ক স্থাপন করে চিন্তা করতে পারে।

## ২. টেস্ট-টাইম কম্পিউট (Test-Time Compute): স্কেলিংয়ের তৃতীয় যুগ

বিগত সাত বছর ধরে মডেলের উন্নতি নির্ভর করত কত বড় মডেলে কত বেশি ডেটা দিয়ে প্রশিক্ষণ দেওয়া হয়েছে তার ওপর। কিন্তু ২০২৪ সালে OpenAI o1 এবং পরবর্তী উদ্ভাবনগুলো প্রমাণ করেছে স্কেলিংয়ের সবচেয়ে শক্তিশালী তৃতীয় স্তম্ভ হলো **টেস্ট-টাইম কম্পিউটেশন**।

কোনো কঠিন অলিম্পিয়াড পর্যায়ের গণিত বা সফটওয়্যার আর্কিটেকচার সমস্যা সমাধান করতে দিলে মডেলটি সরাসরি ঝটপট উত্তর না দিয়ে হাজার হাজার অভ্যন্তরীণ **রিজনিং টোকেন** তৈরি করে গভীরভাবে চিন্তা করে, ভুল হলে পেছনে ফিরে আসে (Backtracking), বিকল্প পথ অনুসন্ধান করে এবং নিজের যুক্তি নিজেই যাচাই করে। ফলে অনেক ছোট মডেলও দীর্ঘ চিন্তা করার সুযোগ পেলে শতগুণ বড় মডেলকে অনায়াসে পরাজিত করতে সক্ষম হয়।

Shakil Mahmud • Research Monograph

## ৩. স্বায়ত্তশাসিত এজেন্ট ও কৃত্রিম প্রজ্ঞার ভবিষ্যৎ

জেনারেটিভ মডেল আজ নিছক চ্যাটবট থেকে পরিণত হচ্ছে সক্রিয় স্বায়ত্তশাসিত এজেন্টে—যা কোড লিখতে পারে, টার্মিনালে চালিয়ে বাগ ঠিক করতে পারে এবং মানুষের হয়ে জটিল গবেষণা সম্পন্ন করতে পারে। ভাষা মডেলের এই যাত্রা কেবল শব্দ গঠনের পরিসংখ্যান নয়; বরং এটি মানব ইতিহাসের সবচেয়ে রোমাঞ্চকর বৈজ্ঞানিক বিপ্লব—সিলিকনের বুকে স্বাধীন কৃত্রিম যুক্তির জন্ম।

*সাইটেশন: মাহমুদ, শাকিল (2026)। "কৃত্রিম প্রজ্ঞার দিগন্ত — মাল্টিমডালিটি, রিজনিং টোকেন ও টেস্ট-টাইম কম্পিউট"। লার্জ ল্যাঙ্গুয়েজ মডেল ও জেনারেটিভ এআই, খণ্ড ২০।*
*https://shakilmahmud.com/bn/large-language-models-part-20-frontiers-synthetic-cognition-multimodal-reasoning-test-time-bn/*

আকর গ্রন্থ ও গবেষণা সূত্র

# নির্বাচিত গ্রন্থপঞ্জি

❖

*এই মনোগ্রাফে আলোচিত তত্ত্ব, গাণিতিক প্রতিপাদন এবং ঐতিহাসিক বিশ্লেষণের ভিত্তি হিসেবে নিম্নলিখিত আকর গ্রন্থ ও প্রাথমিক গবেষণাপত্রসমূহ ব্যবহৃত হয়েছে:*

**Vaswani, A., Shazeer, N., Parmar, N., et al.** (2017). *Attention Is All You Need*. NeurIPS 30: 5998–6008.

**Kaplan, J., McCandlish, S., Henighan, T., et al.** (2020). *Scaling Laws for Neural Language Models*. arXiv:2001.08361 [cs.LG].

**Brown, T. B., Mann, B., Ryder, N., et al.** (2020). *Language Models are Few-Shot Learners*. NeurIPS 33: 1877–1901.

**Dao, T., Fu, D. Y., Ermon, S., Rudra, A., & Ré, C.** (2022). *FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness*. NeurIPS 35: 16344–16359.

**Hu, E. J., Shen, Y., Wallis, P., et al.** (2021). *LoRA: Low-Rank Adaptation of Large Language Models*. ICLR 2022.

**Rafailov, R., Sharma, A., Mitchell, E., et al.** (2023). *Direct Preference Optimization: Your Language Model is Secretly a Reward Model*. NeurIPS 36.

**Song, Y., Sohl-Dickstein, J., Kingma, D. P., et al.** (2020). *Score-Based Generative Modeling through Stochastic Differential Equations*. ICLR 2021.

Shakil Mahmud • Research Monograph

লেখক পরিচিতি

# লেখক পরিচিতি

❖

## Shakil Mahmud

*গবেষক, লেখক ও সফটওয়্যার প্রকৌশলী*

শাকিল মাহমুদ একজন স্বাধীন গবেষক, লেখক ও সফটওয়্যার প্রকৌশলী। তাঁর প্রধান আগ্রহের বিষয় হলো তাত্ত্বিক পদার্থবিজ্ঞান, গণিতের ইতিহাস, আধুনিক ম্যাক্রোঅর্থনীতি, কম্পিউটিংয়ের বিবর্তন এবং স্থান-কালের দার্শনিক তাৎপর্য। তিনি জটিল দার্শনিক ও বৈজ্ঞানিক ধারণাকে গভীর ও সুললিত ভাষায় উপস্থাপন করতে পছন্দ করেন। তাঁর সার্বিক কাজ পাওয়া যাবে তাঁর ব্যক্তিগত গবেষণা পোর্টাল **shakilmahmud.com** এ।

**COLOPHON**

Typeset in EB Garamond (English) and Kohinoor Bangla / Noto Serif Bengali (Bengali).
Mathematical typesetting generated with Donald Knuth's TeX algorithms and KaTeX.
Published digitally as a monolithic research monograph by Shakil Mahmud Editorial Series.
Dhaka & Singapore • 2026 • https://shakilmahmud.com