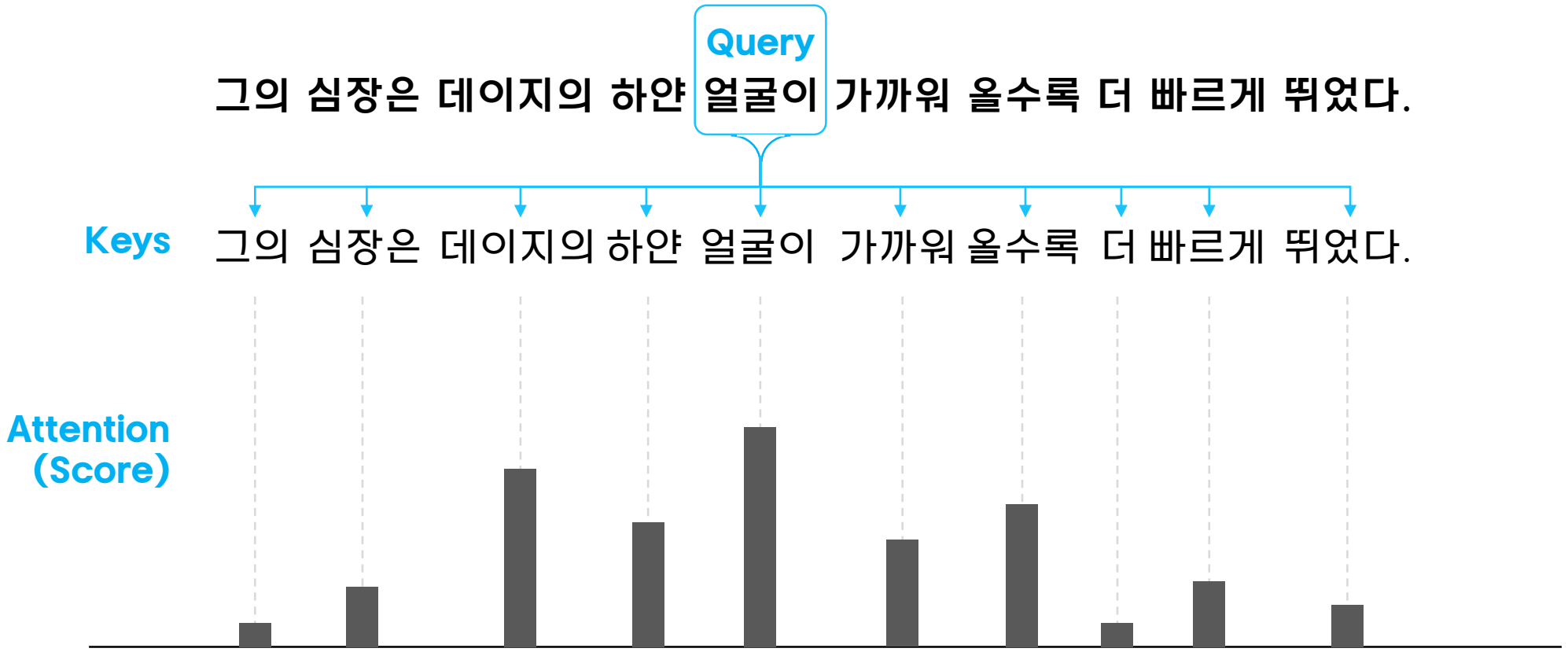
어텐션부터 언어모델로 Agentic AI까지

June 12, 2025

권영준
삼성SDS

어텐션

Mechanism to compute “correlation” for all possible pair of input words



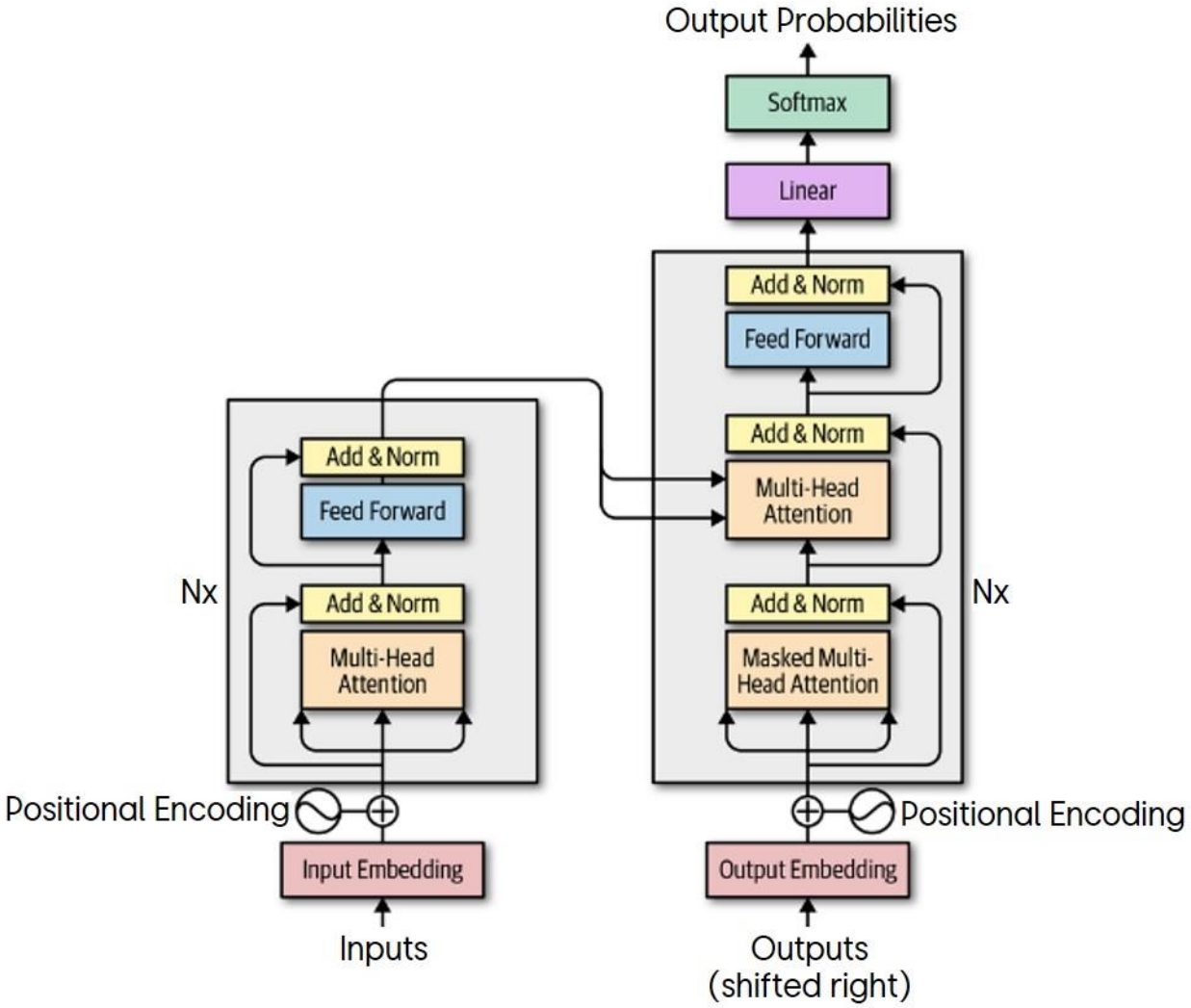
Transformer

Encoder (왼쪽)
Input Embeddings: 토큰을 벡터로 변환
Positional Encoding: 위치 정보 추가
Multi-head Attention: (다각화 된) 어텐션
Add & Norm: 레이어 정규화
Feed Forward: 완전 연결하여 진행

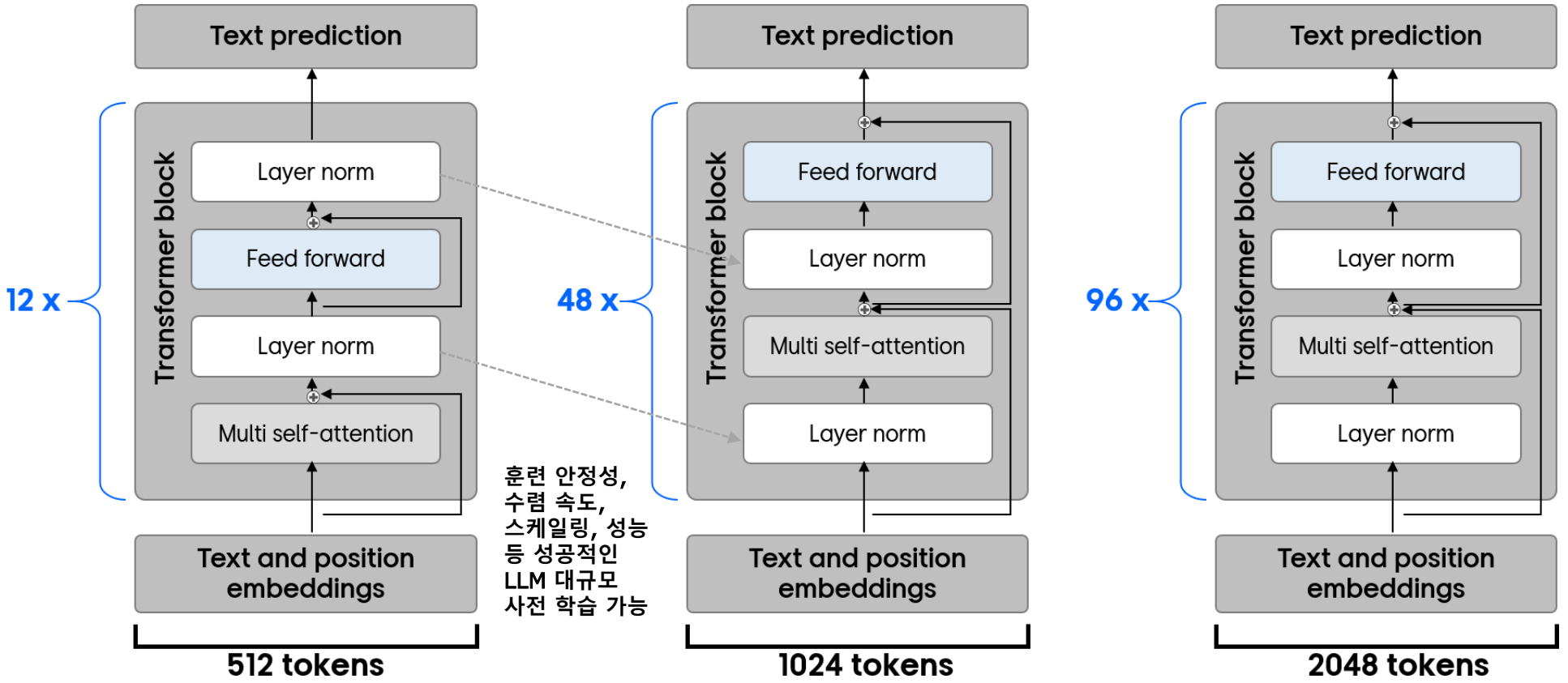
Decoder (오른쪽)
Masked Multi-head Attention: 미래 토큰 마스킹
Cross-attention: 인코더 출력과의 어텐션
Linear & Softmax: 최종 확률 분포 생성

Self-attention: RNN/CNN 없이 순수 어텐션만 사용
Parallel Processing: 순차 처리 없이 병렬 계산 가능
Scalability: 더 긴 시퀀스 처리 가능

현재 GPT, BERT 등 모든 현대 언어 모델의 기반



Scaling Attention Blocks (Transformer) = LLM



GPT-1 (117M) < **GPT-2 (1.5B)** << **GPT-3 (175B)**

* GPT-4: 120x, 1800B

Emergence (창발)

Published is Transactions on Machine Learning Research(08/2022)

Emergent Abilities of Large Language Models

Jason Wei [1] jasonwei@google.com
Yi Tay [1] yitay@google.com
Rishi Bommasani [2] nlprishi@stanford.edu
Colin Raffel [3] craffel@gmail.com
Barret Zoph [1] barretzoph@google.com
Sebastian Borgeaud [4] sborgeaud@deepmind.com
Dani Yogatama [4] dyogatama@deepmind.com
Maarten Bosma [1] bosma@google.com
Denny Zhou [1] dennyzhou@google.com
Donald Metzler [1] metzler@google.com
Ed H. Chi [1] edchi@google.com
Tatsunori Hashimoto [2] thashim@stanford.edu
Oriol Vinyals [4] vinyals@deepmind.com
Percy Liang [2] pliang@stanford.edu
Jeff Dean [1] jeff@google.com
William Fedus [1] liamfedus@google.com

[1] *Google Research* [2] *Stanford University* [3] *UNC Chapel Hill* [4] *DeepMind*

Reviewed on OpenReview: *https://openreview.net/forum?id=yzkSU5zdwD*

Abstract

Scaling up language models has been shown to predictably improve performance and sample efficiency on a wide range of downstream tasks. This paper instead discusses an unpredictable phenomenon that we refer to as *emergent abilities* of large language models. We consider an ability to be emergent if it is not present in smaller models but is present in larger models. Thus, emergent abilities cannot be predicted simply by extrapolating the performance of smaller models. The existence of such emergence raises the question of whether additional scaling could potentially further expand the range of capabilities of language models.

Emergent abilities of LLMs are a pleasant surprise!

Not present in smaller models but present only in large models

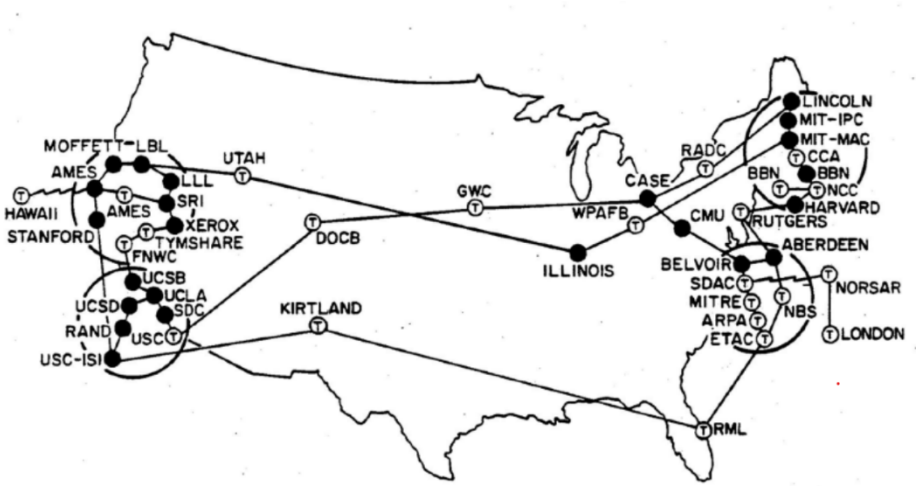
- Prompt engineering, in-context learning
- Zero-shot, few-shot learning
- Logical processing, problem solving
- Reasoning

Emergent abilities couldn't have been guessed

- Can't predict them simply by extrapolating performance of smaller models

Historical Example

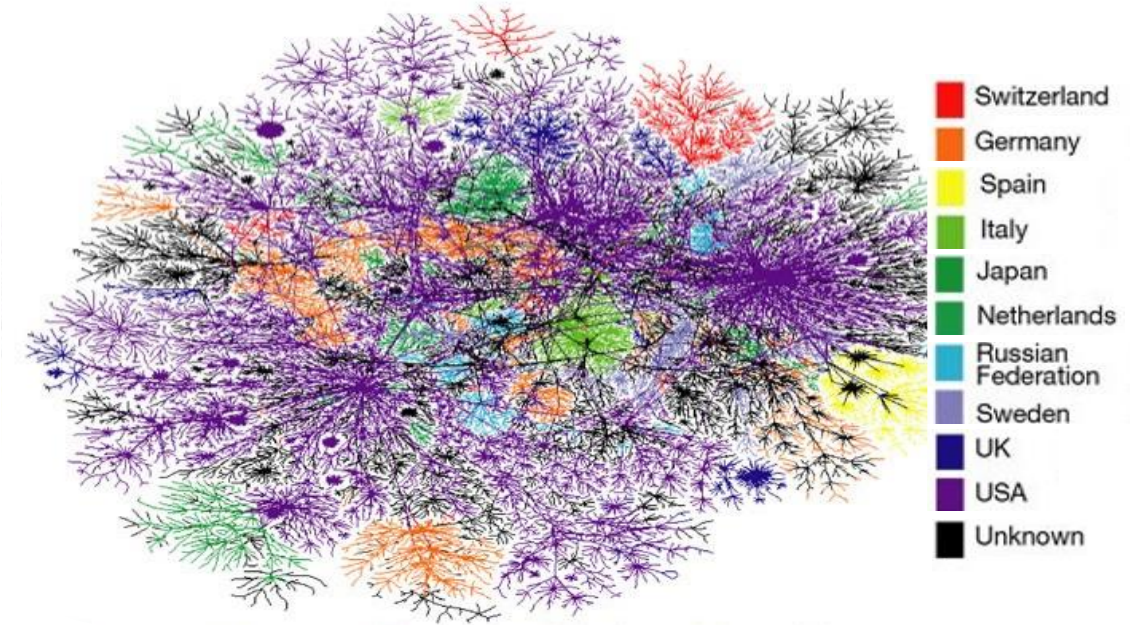
ARPANET (circa 1977)



Simple, Experimental Network

- Limited nodes (4-100)
- Research institutions only
- Basic packet switching
- Proof of concept

Internet 2000



Complex, Emergent Ecosystem

- Billions of connected nodes
- Global digital civilization
- Emergent behaviors
- Transformed humanity

Are Emergent Abilities of Large Language Models a Mirage?

Rylan Schaeffer
Computer Science
Stanford University
rschaef@cs.stanford.edu

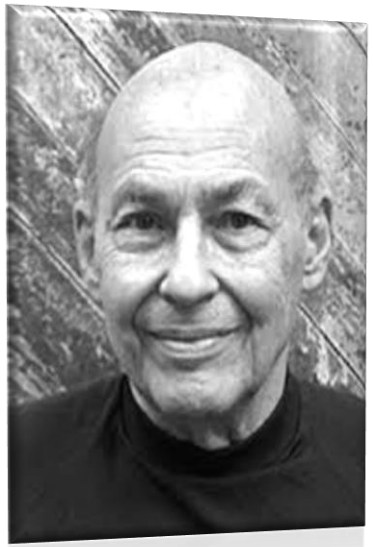
Brando Miranda
Computer Science
Stanford University
brando9@cs.stanford.edu

Sanmi Koyejo
Computer Science
Stanford University
sanmi@cs.stanford.edu

Abstract

Recent work claims that large language models display *emergent abilities*: abilities not present in smaller-scale models that are present in larger-scale models. What makes emergent abilities intriguing is two-fold: their *sharpness*, transitioning seemingly instantaneously from not present to present, and their *unpredictability*, appearing at seemingly unforeseeable model scales. Here, we present an alternative explanation for emergent abilities: for a particular task and model family, when ana-

Minsky 1986

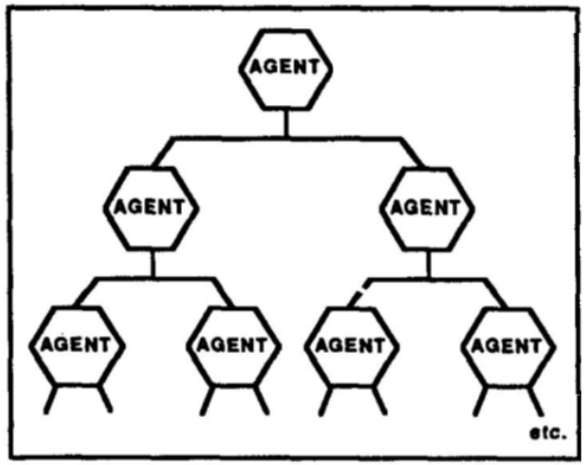


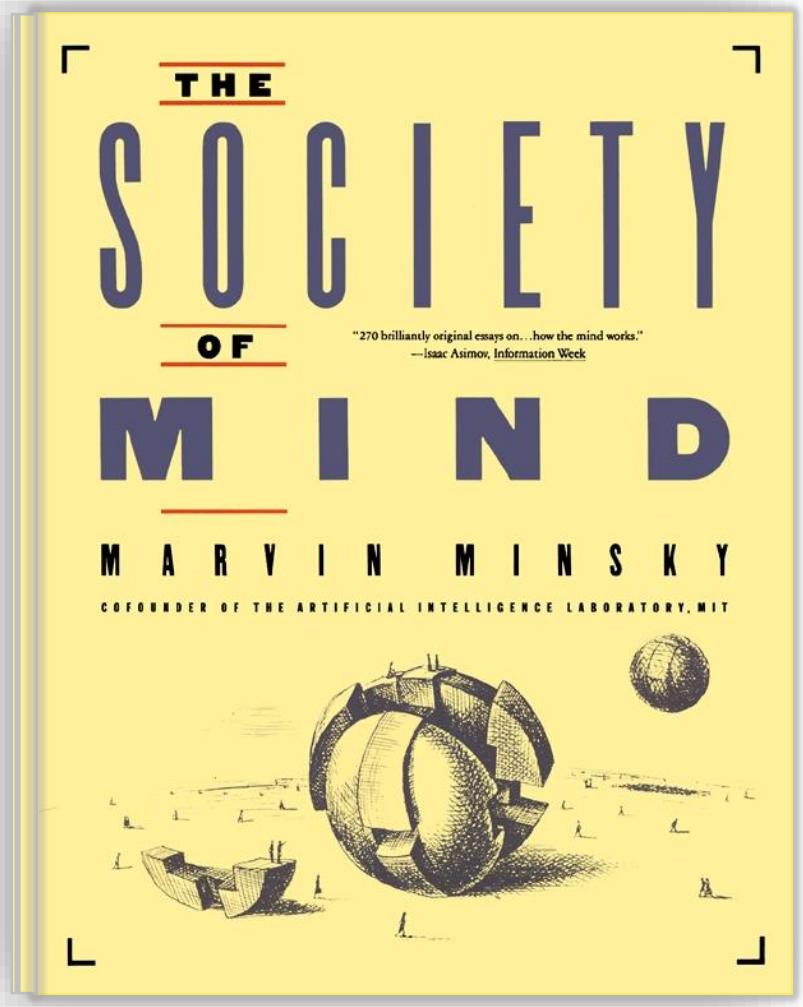
Marvin Minsky
(1927-2016)

Intelligence (**mind**) emerges from many small, specialized processes (**agents**)

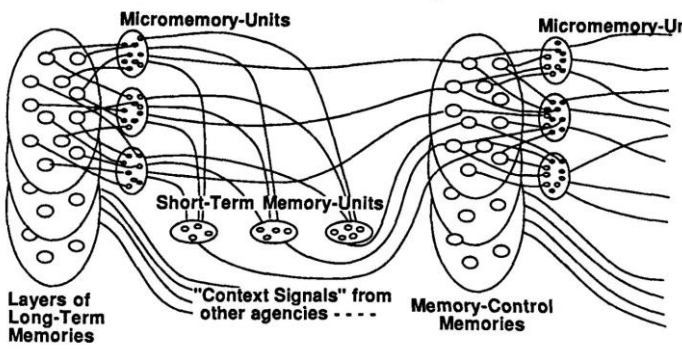
"What magical trick makes us intelligent? The trick is that there is no trick. The power of intelligence stems from our vast diversity, not from any single, perfect principle."

– M. Minsky, 1986

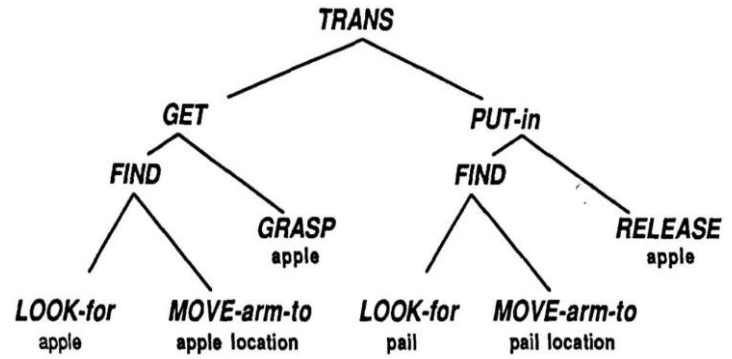




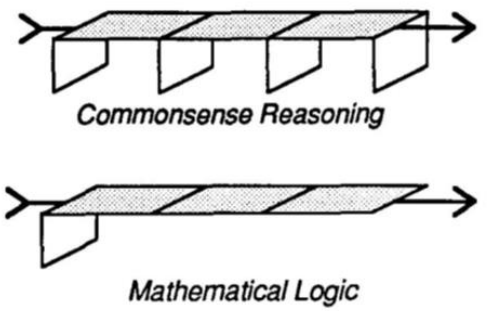
Minsky 1986 (CONT.)



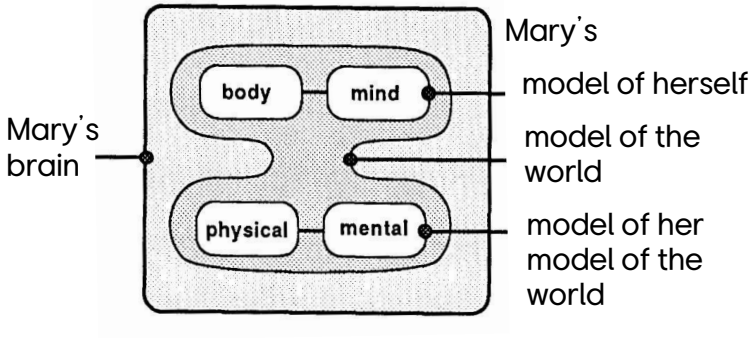
Memory



Inter-agent communication

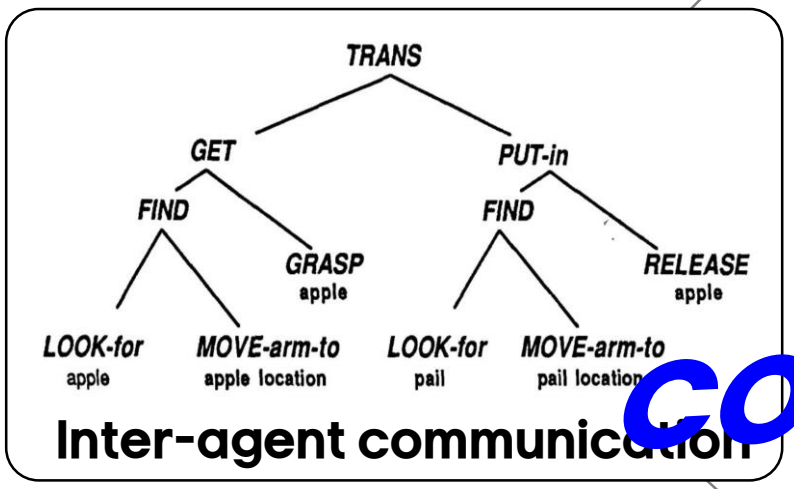


Reasoning, chain of thoughts

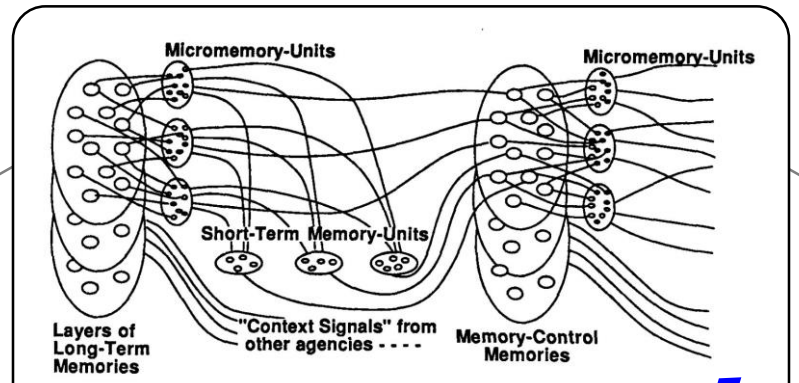


World model

Minsky 1986 (CONT.)

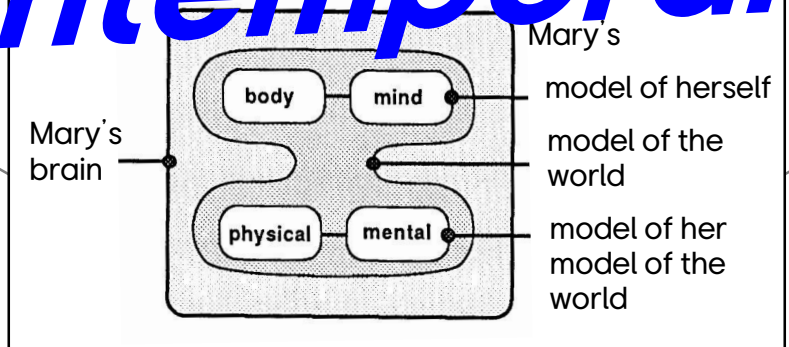


Inter-agent communication

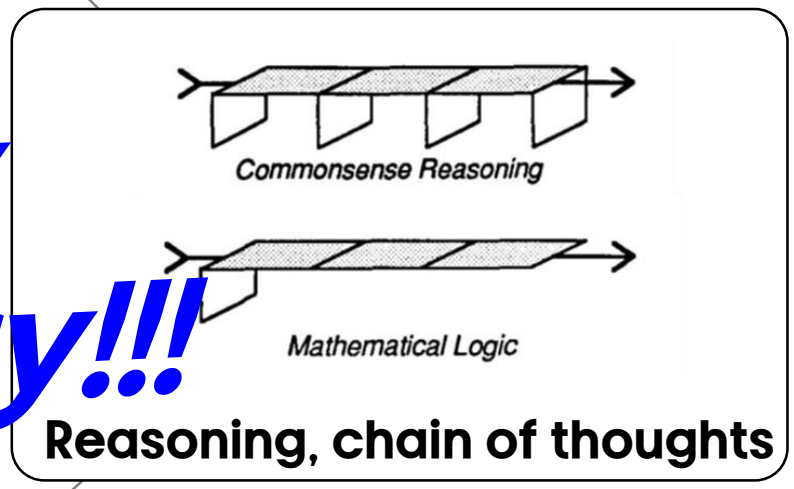


Memory

Shockingly contemporary!!!



World model



Reasoning, chain of thoughts

Agentic AI

OpenAI Deep Research, Operator, Agents SDK

ANTHROP\C Computer Use, Model Context Protocol

Google Deep Research, Agent2Agent Protocol

Microsoft AI Agent Service

Meta Agent Framework

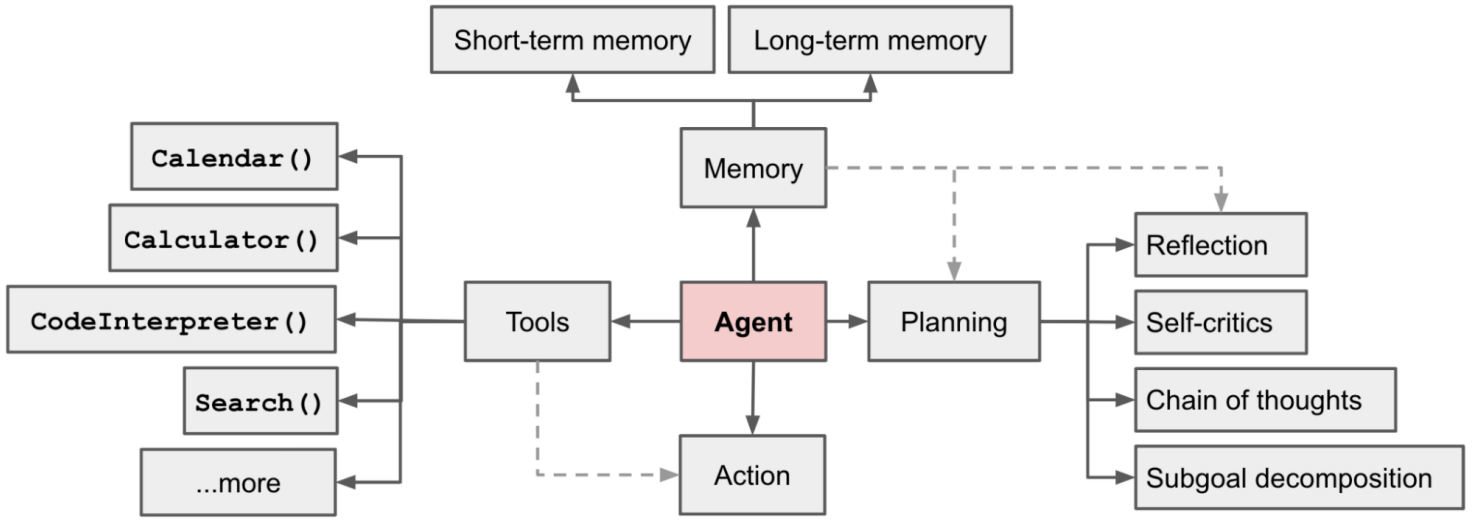
LLM-based

- Language Models as Agent Models [Andreas 2022]

What is missing in LLM?

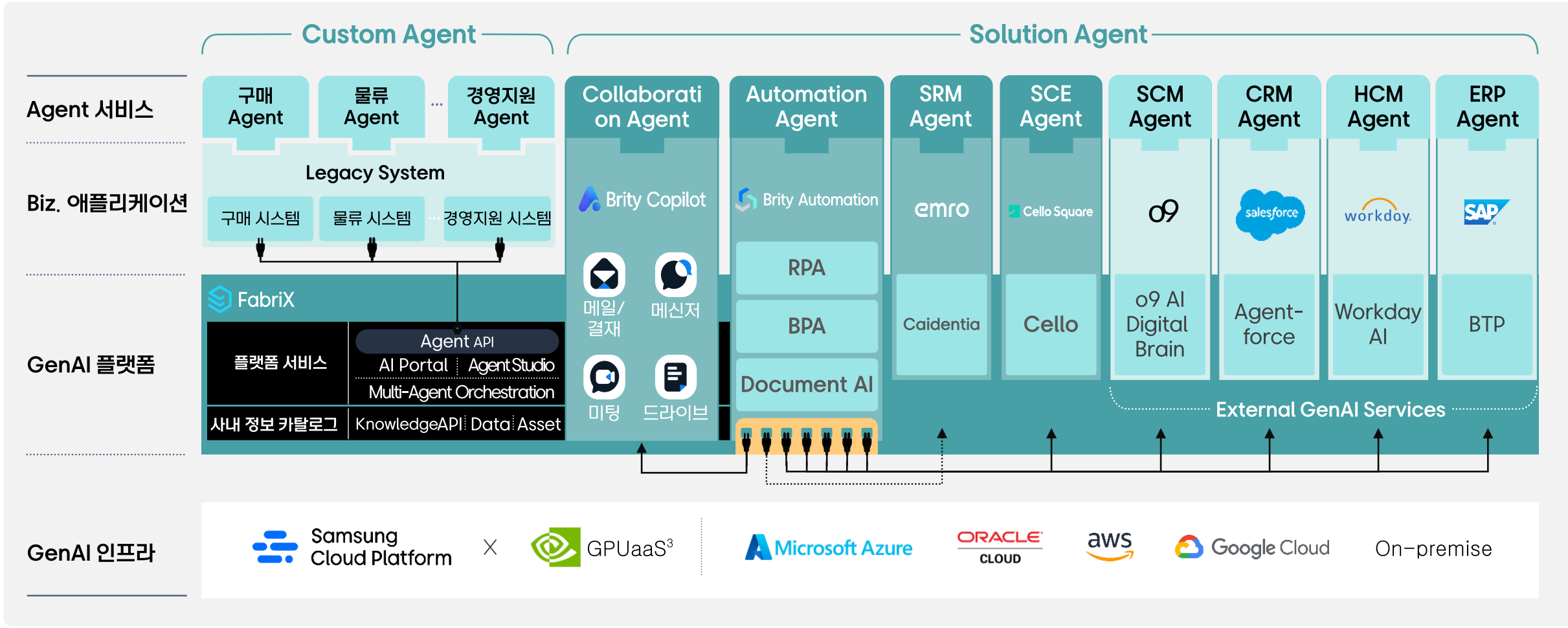
- Communication, memory, learning, embodiment, ...
- Minsky 1986 had already told us about these

LLM Powered Autonomous Agents [Weng 2023]



삼성SDS: Agentic AI for Enterprise

여러 솔루션, 플랫폼 및 시스템을 연결해서 AI 적용 효과와 범위를 극대화



Where Are We Heading?
Emergent Abilities of Agents

Agentic AI as a means for computing with LLMs at scale

Recent breakthroughs

- A world model? Social behavior modeling via multi-agent simulations
- Multi-agent collaboration improves reasoning
- Decision making, tool uses and multiple modalities
- Specialized “Reasoning LLMs”

Caution Required

⚠️ Must be well coordinated and governed

⚠️ Autonomous decision making is dangerous

⚠️ Safety, safety, safety!

⚠️ Security, security, security!

- Still, I’m expecting a bigger surprise; A much bigger, much more pleasant surprise from computing agentic AI at scale than a doomsday for earth!

감사합니다!