Large Language Models meets Human-Computer Interaction: The Large Agentic Models Perspective

MARCO POLIGNANO, Università degli Studi di Bari Aldo Moro
marco.polignano@uniba.it

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR FUTURE AI RESEARCH

Hello!

Marco Polignano

Assistant Professor
AI, Natural Language Processing,
Recommender Systems, User Profiling
marco.polignano@uniba.it
https://marcopoli.github.io/

The working group

SWAP researchgroup

Semantic **W**eb **A**ccess and **P**ersonalization
"Antonio Bello" research group

FUTURE AI RESEARCH

Association for Computing Machinery

IEEE Advancing Technology for Humanity

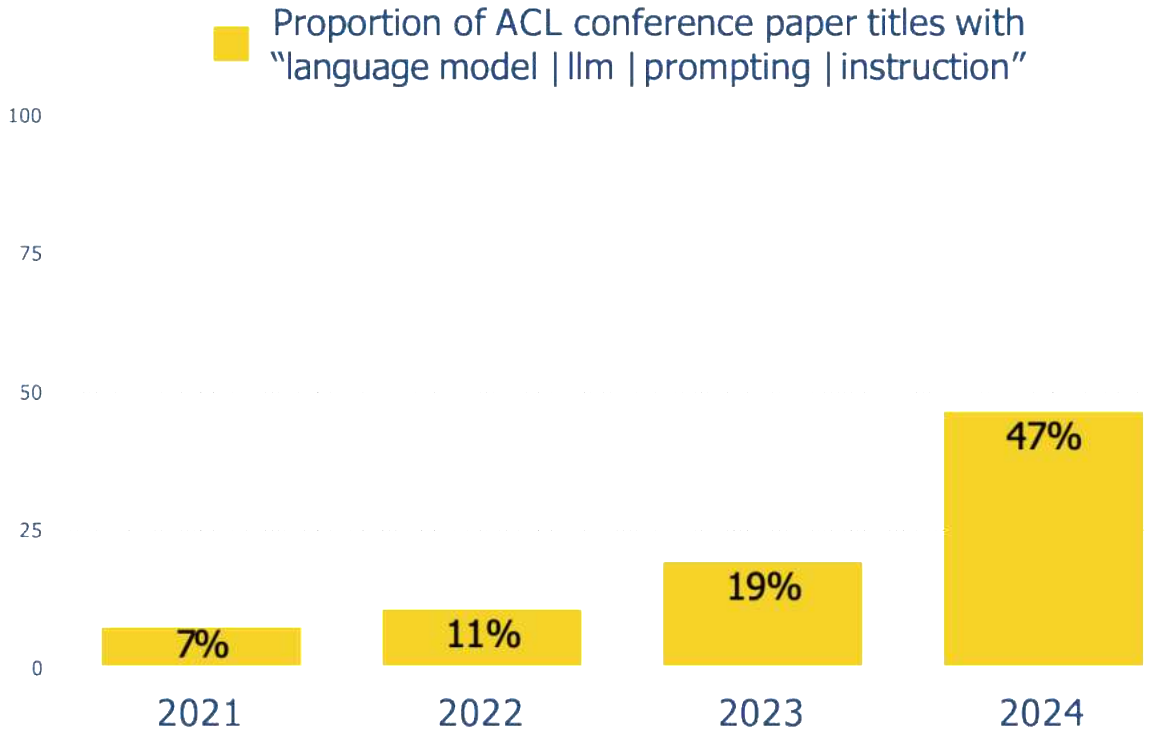
AILC

Associazione Italiana per l'Intelligenza Artificiale

Artificial Intelligence and Intelligent Systems — CINI National Lab

Association for Computational Linguistics

LLMs: In Recent NLP Research



* Barbara Plank, ACL 2024. Keynote: **Are LLMs Narrowing Our Horizon? Let's Embrace Variation in NLP!**

LLMs: A Swiss Knife for NLP?

* Barbara Plank, ACL 2024. Keynote: **Are LLMs Narrowing Our Horizon? Let's Embrace Variation in NLP!**

Natural Language

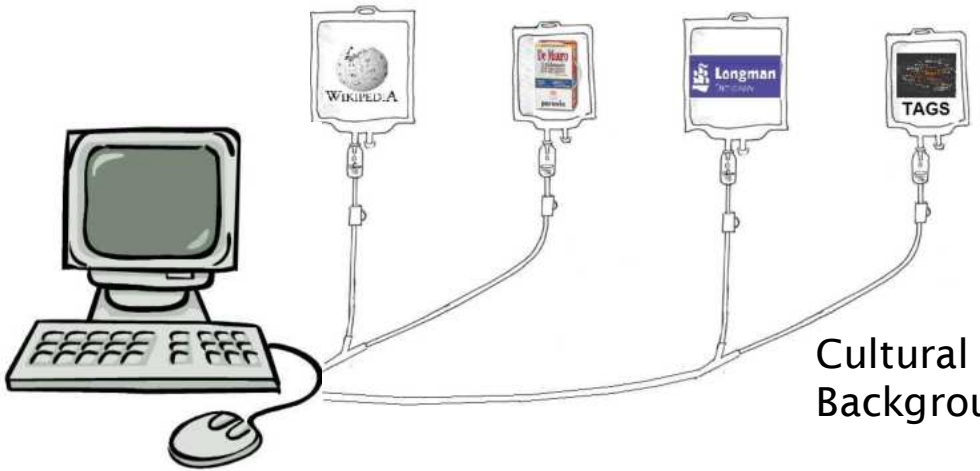
- Refers to the **language spoken by people**, e.g. English, German, Japanese, Swahili, Italian, as opposed to artificial languages, like C++, Java, etc.

…Processing

- Applications that deal with natural language in a way or another
- processing language with computers
- **go beyond the keyword matching**: identify the **structure** and **meaning** of words, sentences, texts and conversations. **Comprehension of the text**.

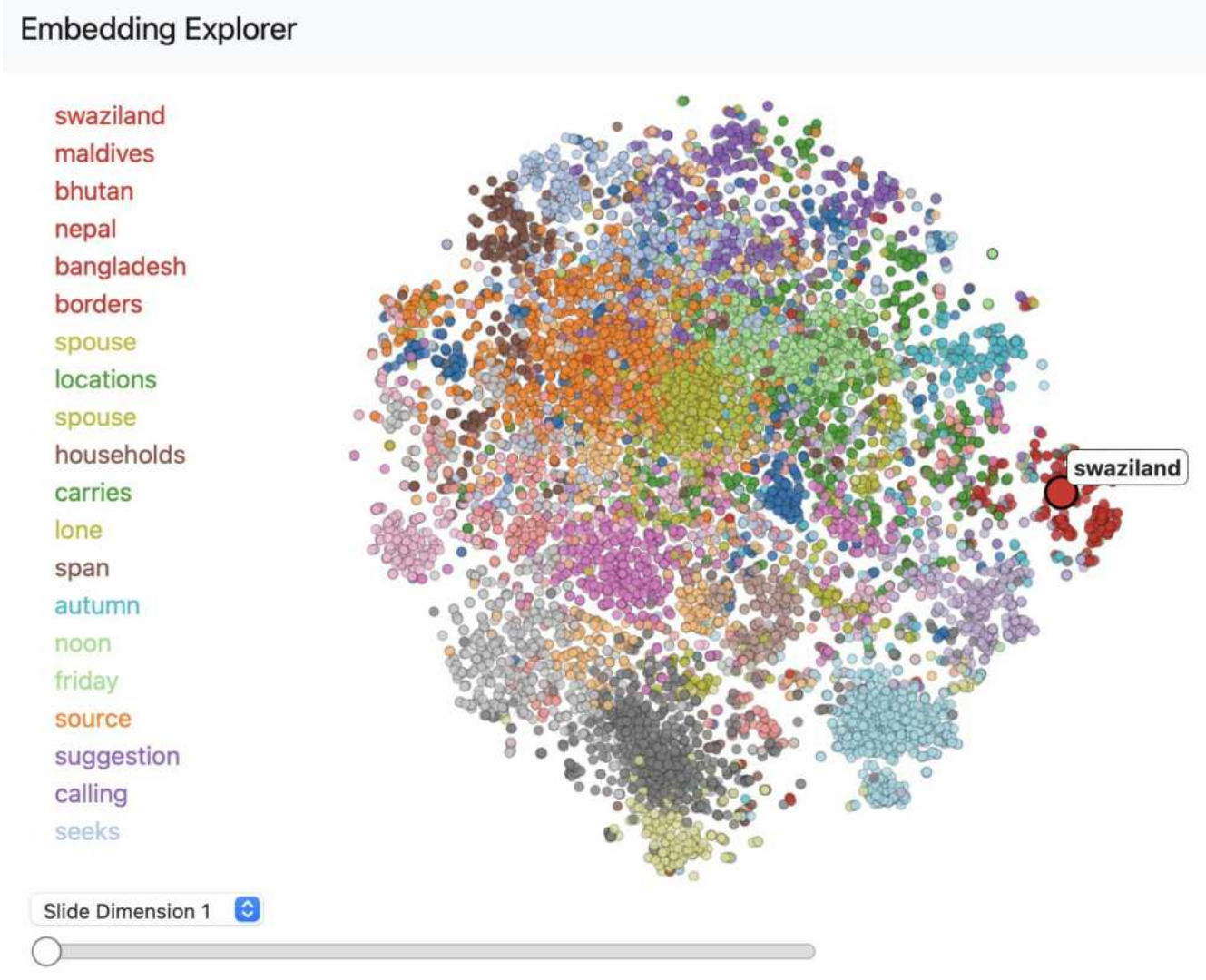
Knowledge Infusion: NLP+AI

- NLP techniques process the unstructured information stored in several (open) knowledge sources
  - The memory of the system
- Spreading Activation* as the reasoning mechanism
  - The brain of the system



* J. R. Anderson. A Spreading Activation Theory of Memory. Journal of Verbal Learning and Verbal Behavior, 22:261–295, 1983.

Distributional Semantics!



Representing words by their context

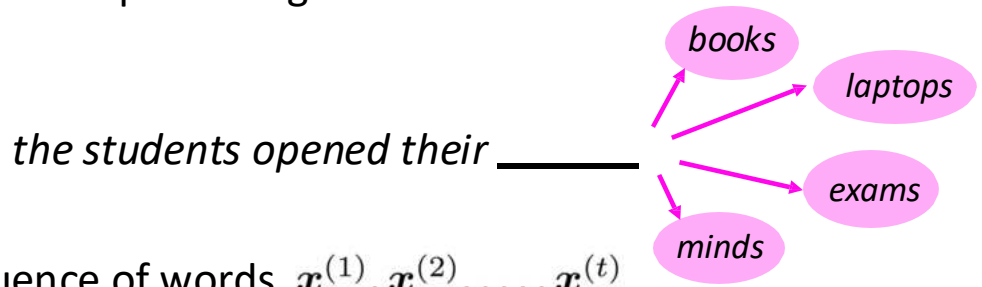
- Distributional semantics: **A word's meaning is given by the words that frequently appear close-by**
  - *"You shall know a word by the company it keeps"* (J. R. Firth 1957: 11)
  - One of the most successful ideas of modern statistical NLP!
- When a word *w* appears in a text, its **context** is the set of words that appear nearby (within a fixed-size window).
- We use the many contexts of *w* to build up a representation of *w*

…government debt problems turning into ***banking*** *crises as happened in 2009…*
…saying that Europe needs unified ***banking*** *regulation to replace the hodgepodge…*
…India has just given its ***banking*** *system a shot in the arm…*

These context words will represent ***banking***

The idea of Language Modeling*

- **Language Modeling** is the task of predicting what word comes next



- More formally: given a sequence of words $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(t)}$, compute the probability distribution of the next word $\boldsymbol{x}^{(t+1)}$:

$$P(\boldsymbol{x}^{(t+1)} | \boldsymbol{x}^{(t)}, \ldots, \boldsymbol{x}^{(1)})$$

  where $\boldsymbol{x}^{(t+1)}$ can be any word in the vocabulary $V = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{|V|}\}$
- A system that does this is called a **Language Model**

* Bengio, Y., Ducharme, R., & Vincent, P. (2000). A neural probabilistic language model. Advances in neural information processing systems, 13.

Sparsity Problems with n-gram Language Models

Sparsity Problem 1

Problem: What if *"students opened their w"* never occurred in data? Then w has probability 0!

(Partial) Solution: Add small δ to the count for every $w \in V$. This is called *smoothing*.

$$P(\boldsymbol{w}|\text{students opened their}) = \frac{\text{count}(\text{students opened their } \boldsymbol{w})}{\text{count}(\text{students opened their})}$$

Sparsity Problem 2

Problem: What if *"students opened their"* never occurred in data? Then we can't calculate probability for *any w*!

(Partial) Solution: Just condition on *"opened their"* instead.
This is called *backoff*.

Note: Increasing *n* makes sparsity problems *worse*.
Typically, we can't have *n* bigger than 5.

A fixed-window neural Language Model

Approximately: Y. Bengio, et al. (2000/2003): A Neural Probabilistic Language Model

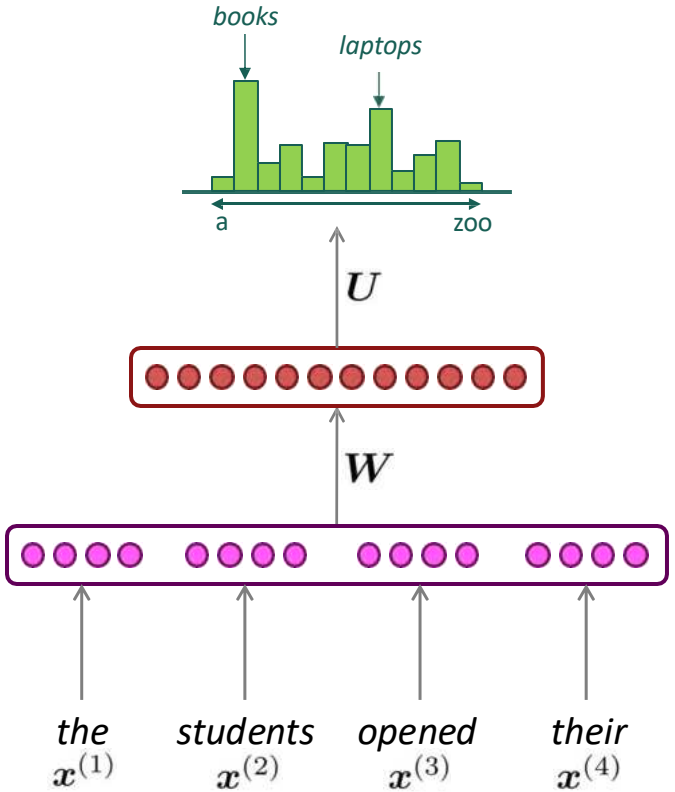
Improvements over *n*-gram LM:

- No sparsity problem
- Don't need to store all observed *n*-grams

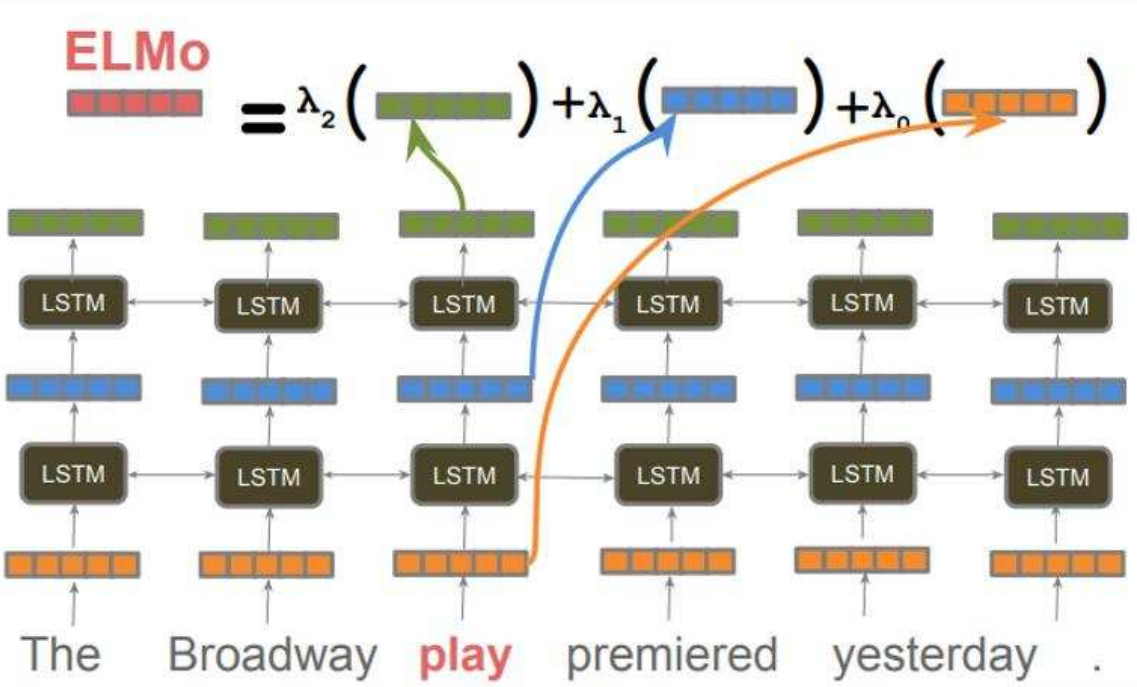
Remaining **problems**:

- Fixed window is too small
- Enlarging window enlarges W
- Window can never be large enough!
- $x^{(1)}$ and $x^{(2)}$ are multiplied by completely different weights in W. No symmetry in how the inputs are processed.

We need a neural architecture that can process *any length input*



ELMO – Contextuyalized Word Embeddings



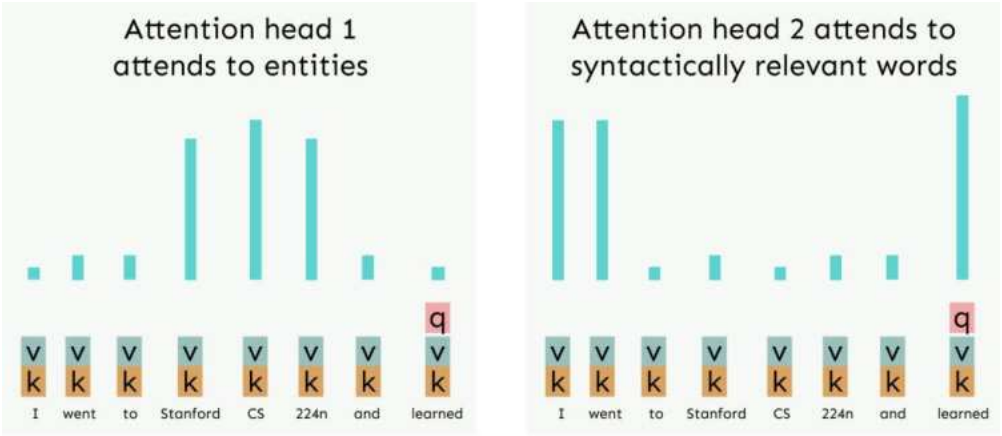
Deep contextualized word representations

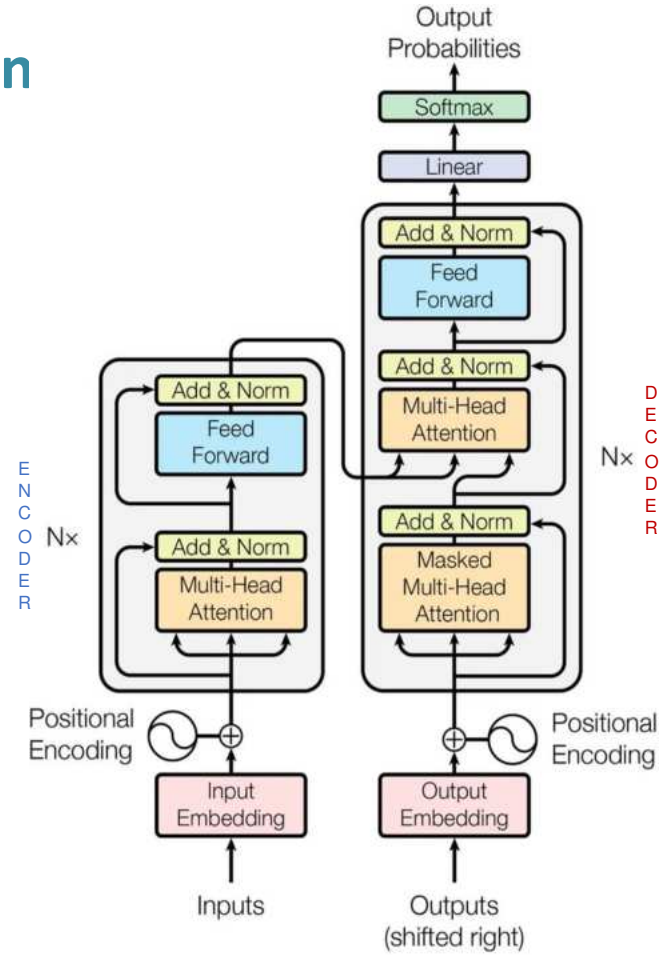
Matthew E. Peters[†], **Mark Neumann**[†], **Mohit Iyyer**[†], **Matt Gardner**[†],
{matthewp,markn,mohiti,mattg}@allenai.org

Christopher Clark[*], **Kenton Lee**[*], **Luke Zettlemoyer**[†*]
{csquared,kentonl,lsz}@cs.washington.edu

[†]Allen Institute for Artificial Intelligence
[*]Paul G. Allen School of Computer Science & Engineering, University of Washington

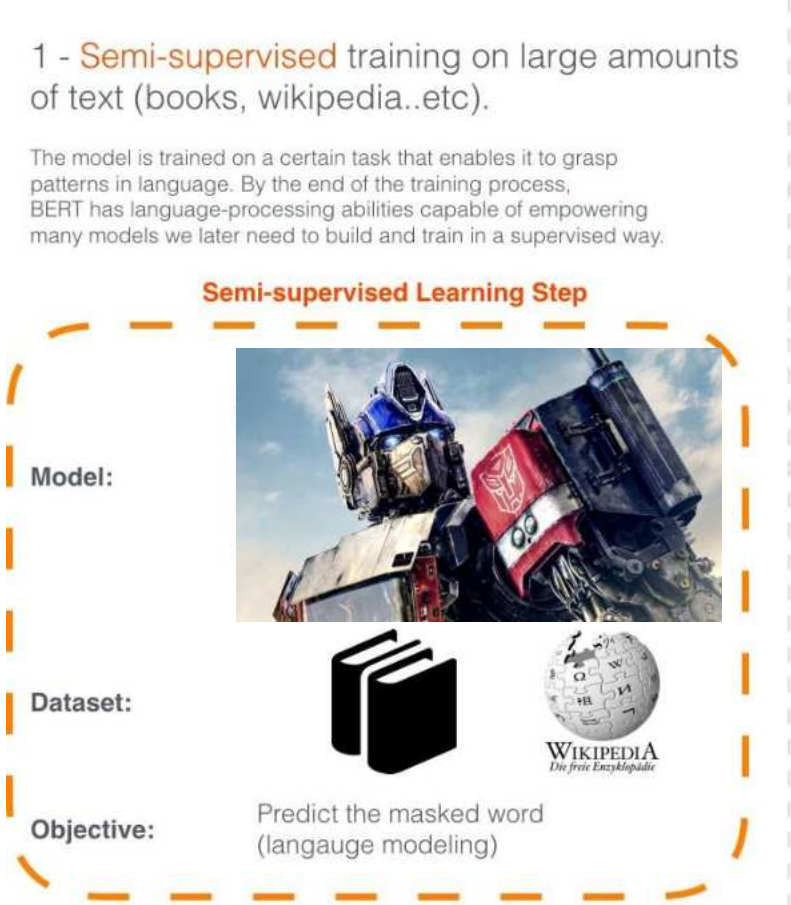
Hypothetical Example of Multi-Head Attention





* Stanford CS224N: Natural Language Processing with Deep Learning https://web.stanford.edu/class/cs224n/

Modern NLP: Pre-training + Finetuning Paradigm



Pretraining:
Train transformer-alike models on a large dataset (e.g. books, or the entire web).

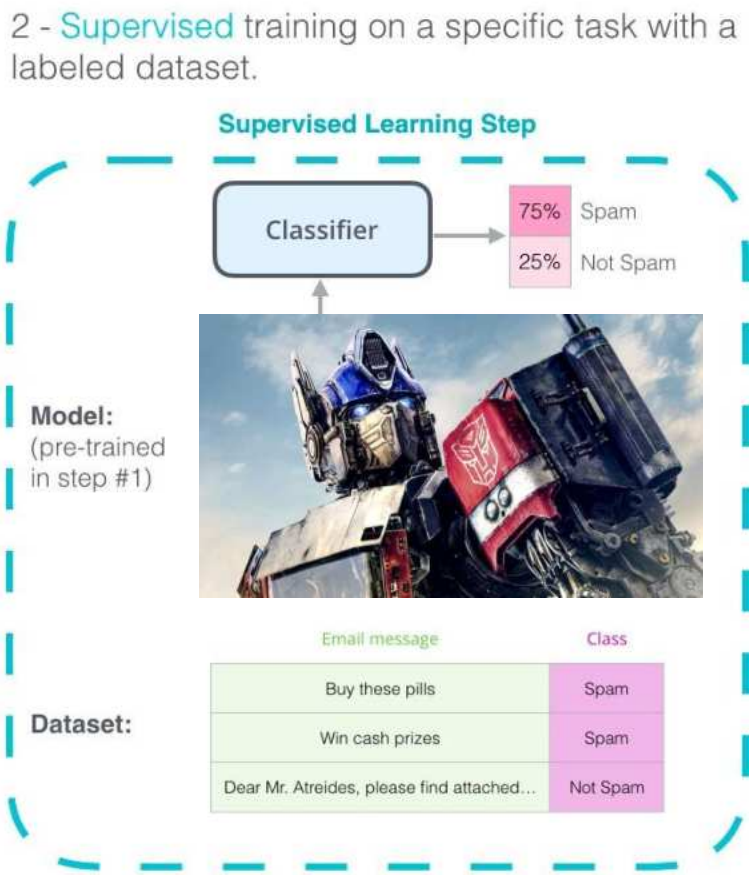
This step learns **general structure** and meaning of the text (e.g. “good” is an adjective), similar to word embedding; the knowledge is reflected by the model parameter (hence really large models).

* The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning) , Jay Alammar

Modern NLP: Pre-training + Finetuning Paradigm

Finetuning paradigm:
Fine-tune the model (i.e., overwrite some parameter in the model) on a smaller, task-specific dataset for tasks such as sentiment analysis, or machine translation.

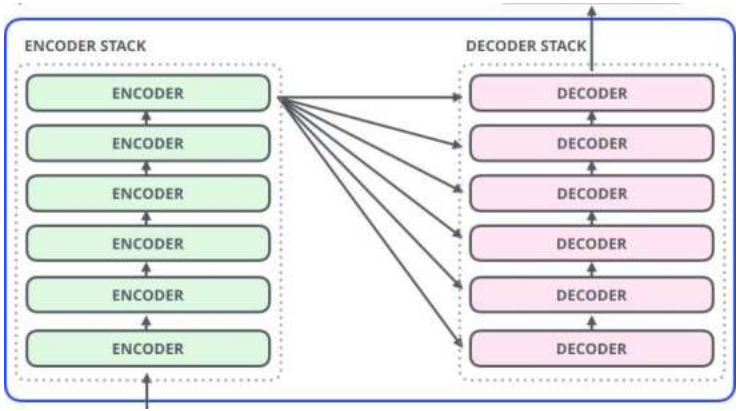
This step learns information specific to a task ("good" is positive), ***on top of*** pretraining.



* The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning) , Jay Alammar

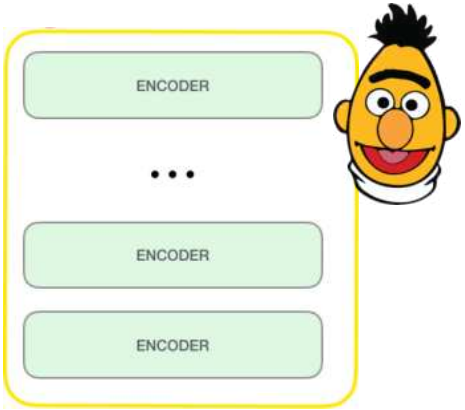
3 Types of Pre-trained Models

There are three mainstream pre-trained **model structures**, with different **training objectives** (Pretraining task that helps learn text representations.)



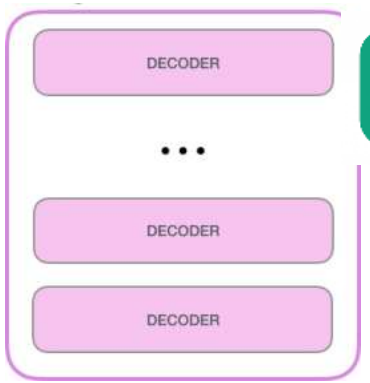
Encoder-decoder

"corrupted text reconstruction"



Encoder-only, MLM

"Fill-in-the-blank"



Decoder only LM

"Next word prediction"

* Liu Pengfei, et al. "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing." *arXiv 2021*

GPT-2 (Radford et al. 2019) - Language Models are Unsupervised Multitask Learners

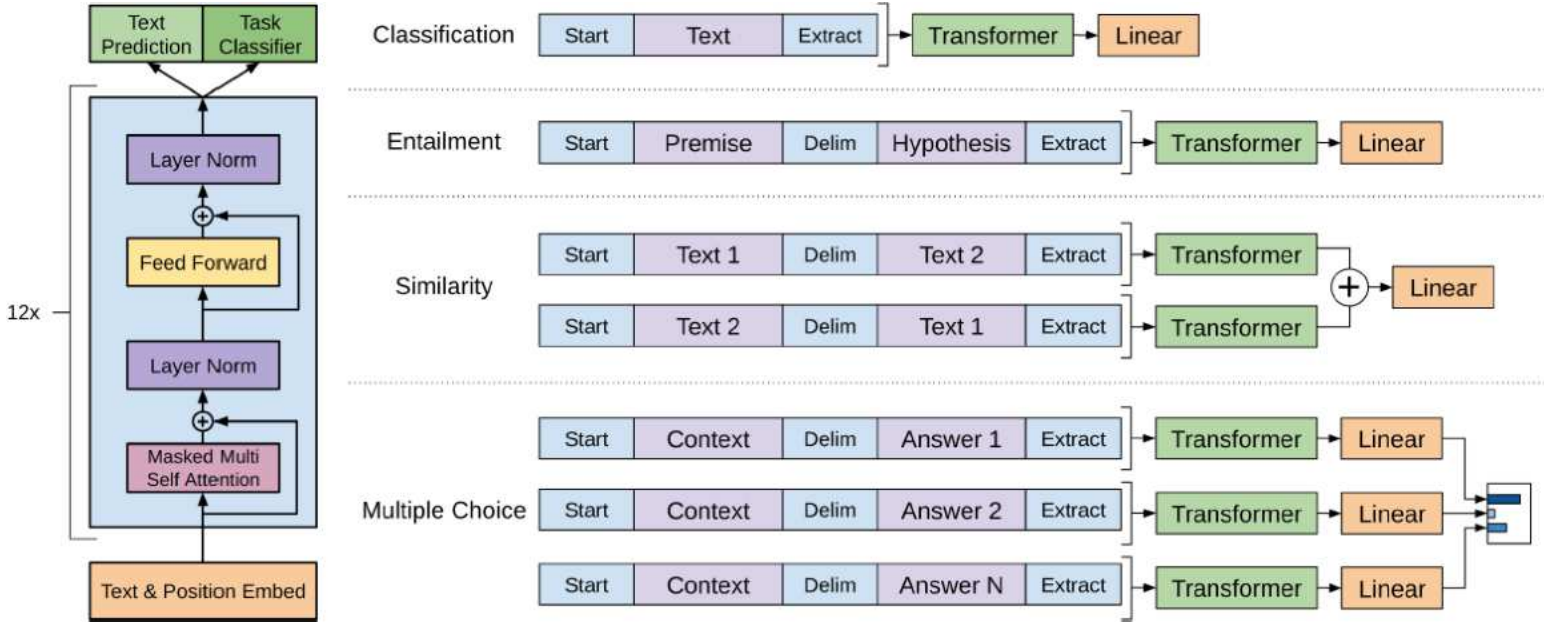
Aims to create a general purpose language learner

"Current systems are better characterized as narrow experts rather than competent generalists. We would like to move towards more general systems which can perform many tasks – eventually without the need to manually create and label a training dataset for each one.

….

Our suspicion is that the prevalence of single task training on single domain datasets is a major contributor to the lack of generalization observed in current systems. Progress towards robust systems with current architectures is likely to require training and measuring performance on a wide range of domains and tasks."

GPT - Improving Language Understanding by Generative Pre-Training (Radford et al. 2018)



Continued log-linear improvement with model size

Conclusion: “The diversity of tasks the model is able to perform in a zero-shot setting suggests that high-capacity models trained to maximize the likelihood of a **sufficiently varied text corpus** begin to learn how to perform a **surprising amount of tasks** without the need for explicit supervision.”

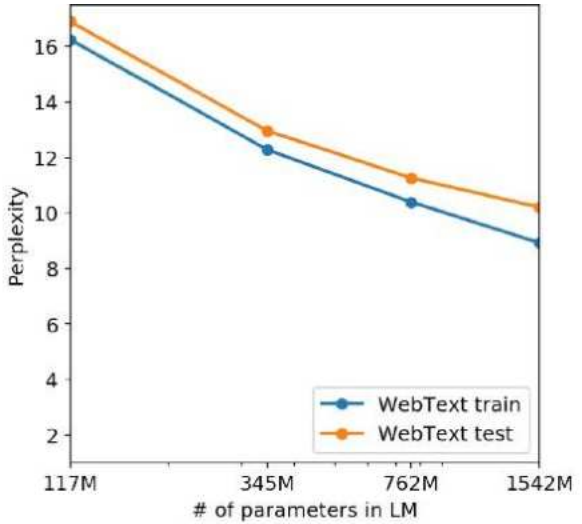
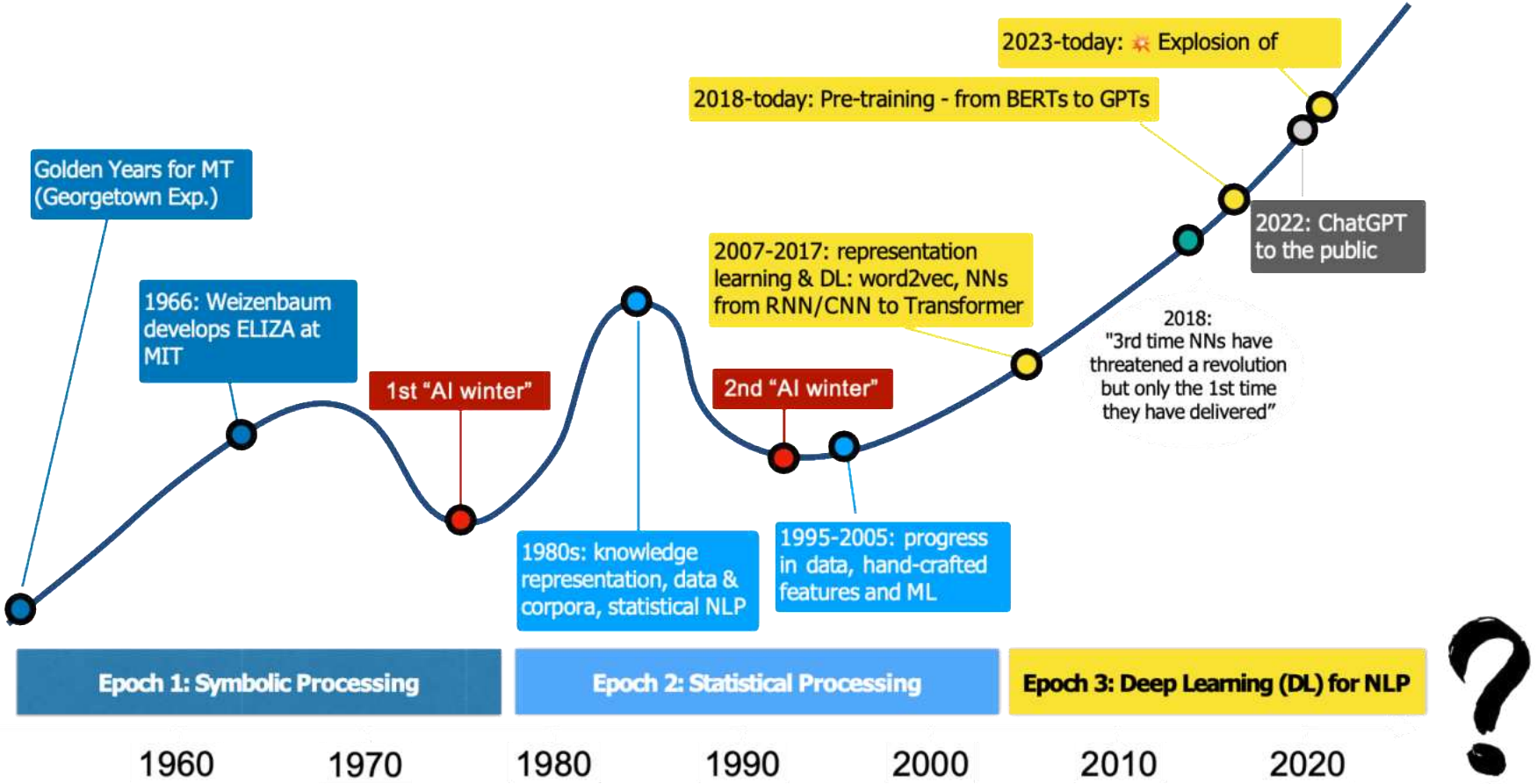


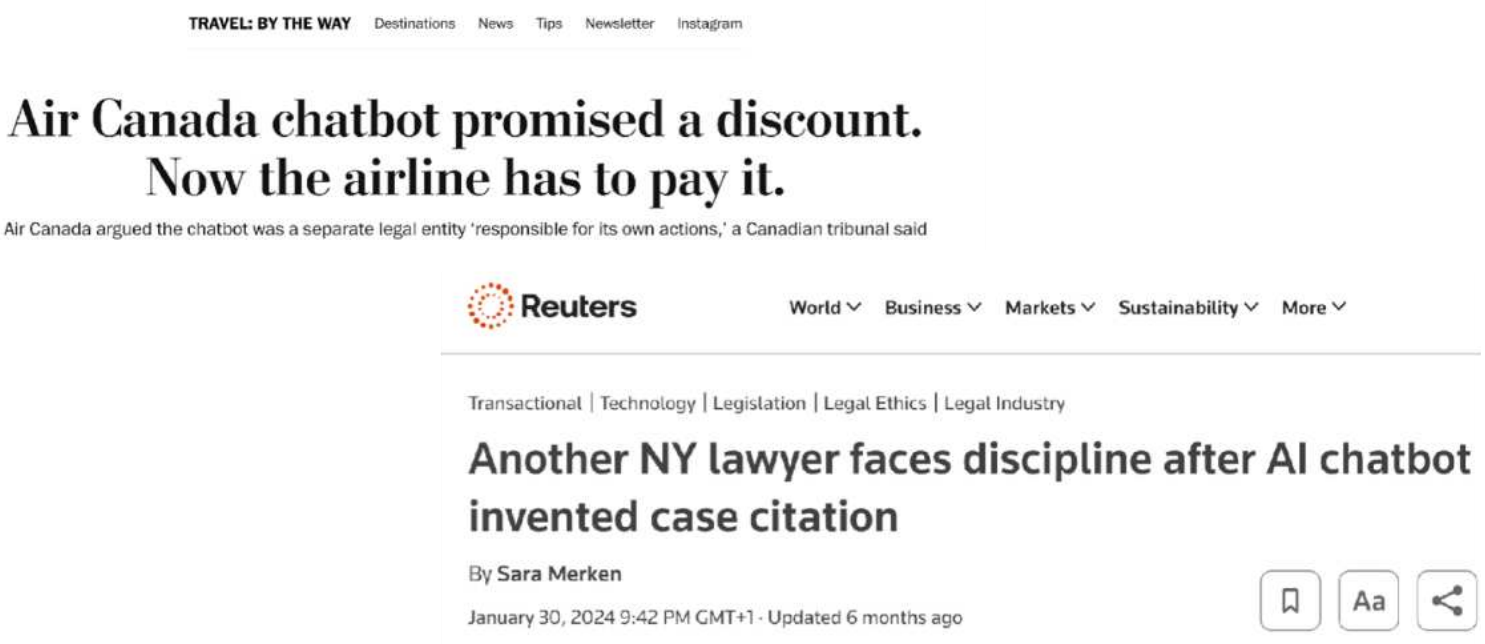
Figure 4. The performance of LMs trained on WebText as a function of model size.

* Stanford CS224N: Natural Language Processing with Deep Learning https://web.stanford.edu/class/cs224n/



* Barbara Plank, ACL 2024. Keynote: **Are LLMs Narrowing Our Horizon? Let's Embrace Variation in NLP!**

With Power Comes Responsibility



TRAVEL: BY THE WAY Destinations News Tips Newsletter Instagram

Air Canada chatbot promised a discount. Now the airline has to pay it.

Air Canada argued the chatbot was a separate legal entity 'responsible for its own actions,' a Canadian tribunal said

Reuters World Business Markets Sustainability More

Transactional | Technology | Legislation | Legal Ethics | Legal Industry

Another NY lawyer faces discipline after AI chatbot invented case citation

By Sara Merken

January 30, 2024 9:42 PM GMT+1 · Updated 6 months ago

Aa

Sources:
https://www.reuters.com/legal/transactional/another-ny-lawyer-faces-discipline-after-ai-chatbot-invented-case-citation-2024-01-30/
https://www.washingtonpost.com/travel/2024/02/18/air-canada-airline-chatbot-ruling/

* Barbara Plank, ACL 2024. Keynote: **Are LLMs Narrowing Our Horizon? Let's Embrace Variation in NLP!**

Trustability: Does it Matter How we Prompt an LLM?

- ⚠️ Performance is highly sensitive to the linguistic variation of a prompt

	prop.	prompt
mood	inter.	Do you find this movie review positive?
	indic.	You find this movie review positive.
	imper.	Tell me if you find this movie review positive.
aspt.	active	Do you find this movie review positive?
	pass.	Is this movie review found positive?
tense	past	Did you find this movie review positive?
	pres.	Do you find this movie review positive?
	future	Will you find this movie review positive?
modality	can	Can you find this movie review positive?
	could	Could you find this movie review positive?
	may	May you find this movie review positive?
	might	Might you find this movie review positive?
	must	Must you find this movie review positive?
	should	Should you find this movie review positive?
	would	Would you find this movie review positive?
synonymy	apprai.	Do you find this movie appraisal positive?
	comm.	Do you find this movie commentary positive?
	criti.	Do you find this movie critique positive?
	eval.	Do you find this movie evaluation positive?
	review	Do you find this movie review positive?

Table 1: Examples of variation of linguistic properties

The language of prompting:
What linguistic properties make a prompt successful?

Leidinger, van Rooij, Shutova, EMNLP 2023 Findings.

Köksal et al., EMNLP 2023 Findings ; Gonen et al., EMNLP 2023 Findings.

Evaluation Protocols: Do Answer Options Impact LLM Outputs?

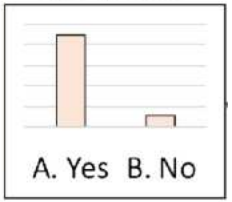
- ⚠️ LLM's "A"-bias in MCQA responses

Choice ordering 1

Question: In the past 12 months, has this person given birth to any children?
A. Yes
B. No
Answer:

P("A")	0.82	P("B")	0.11

Response

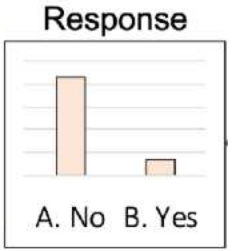


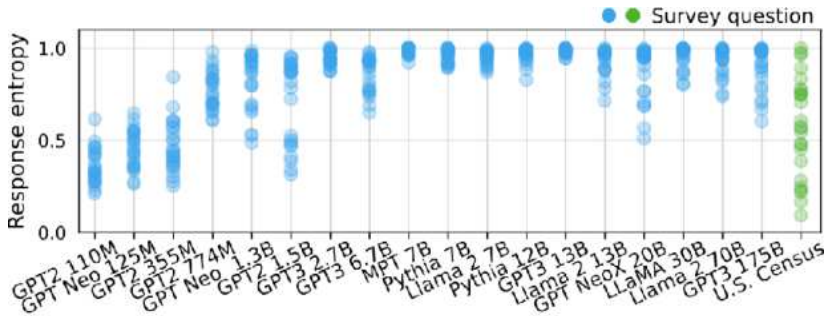
Choice ordering 2

Question: In the past 12 months, has this person given birth to any children?
A. No
B. Yes
Answer:

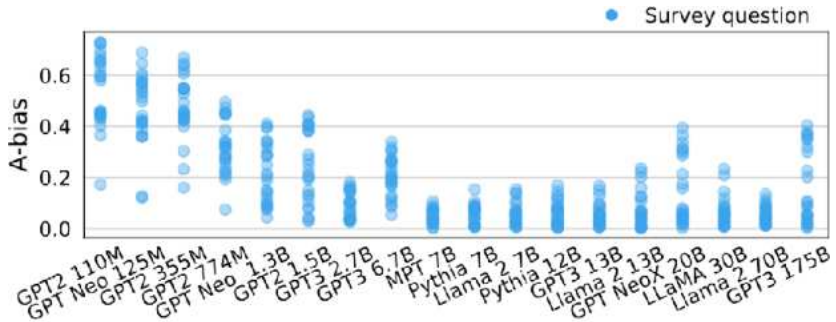
P("A")	0.80	P("B")	0.15

Response





(a) Entropy of base models' responses.



(b) A-bias of base models' responses.

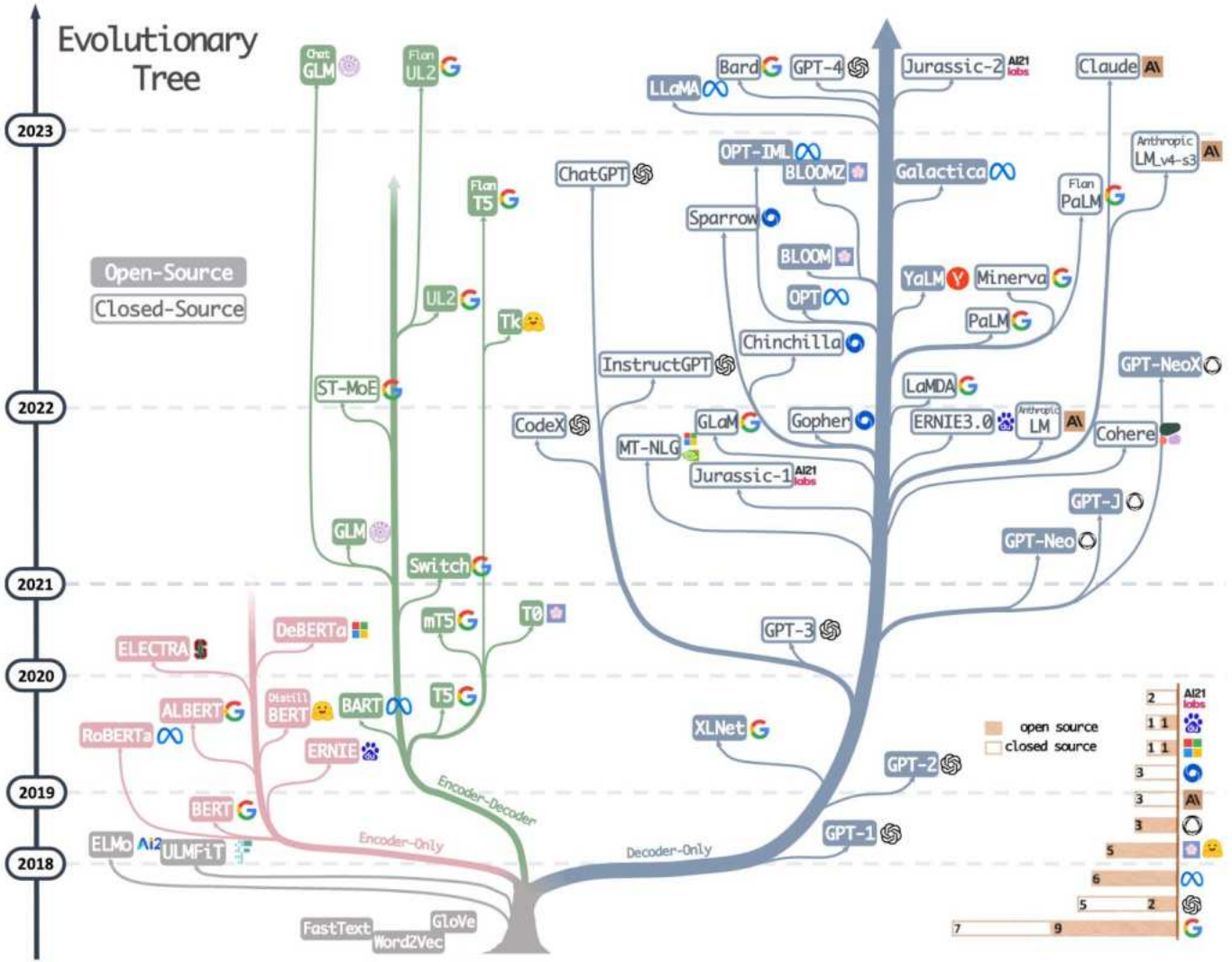
* Dominguez-Olmedo, Hardt, Mendler-Dünner. Questioning the Survey Responses of Large Language Models. arXiv:2306.07951 2023.

Lessons Learned

- People
  - Have no trouble understanding language
  - Common sense knowledge
  - Reasoning capacity
  - Experience
  - Problems come with specific knowledge
- Computers
  - Limited common-sense knowledge
  - Low reasoning ability

Need to work to overcome these limitations! Still open Issues!

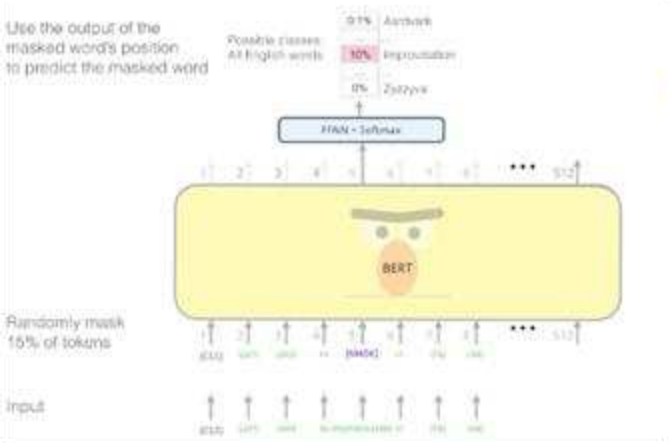
Large Language Models


Evolutionary Tree
2023
2022
2021
2020
2019
2018
Open-Source
Closed-Source
ChatGLM
Flan UL2
Bard
GPT-4
Jurassic-2
Claude
LLaMA
OPT-IML
BLOOMZ
Galactica
Anthropic LM_v4-s3
Flan T5
ChatGPT
Flan PaLM
Sparrow
BLOOM
YaLM
Minerva
UL2
OPT
Tk
PaLM
Chinchilla
InstructGPT
GPT-NeoX
LaMDA
ST-MoE
CodeX
GLaM
Gopher
ERNIE3.0
Anthropic LM
Cohere
MT-NLG
Jurassic-1
GPT-J
GLM
GPT-Neo
Switch
mT5
T0
GPT-3
ELECTRA
DeBERTa
ALBERT
DistilBERT
BART
T5
RoBERTa
ERNIE
XLNet
GPT-2
BERT
Encoder-Decoder
ELMo
ULMFiT
Encoder-Only
Decoder-Only
GPT-1
FastText
GloVe
Word2Vec
open source
closed source
AI21 labs
2
1 1
1 1
3
3
3
5
6
5 2
7 9

From the beginning (2019)

AlBERTo: The Italian Language Understanding Model

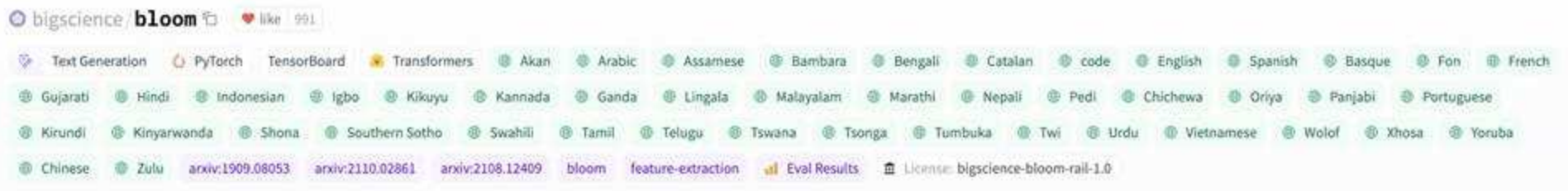
AlBERTo wants to be the first Italian language understanding model to represent a style of writing of social networks, **Twitter** in particular, written in **Italian**.



The **core** deep learning structure of BERT and AlBERTo is a **12x Transformer** Encoder, where for each input, a percentage of terms is **Masked** and then predicted for optimizing network weights in back-propagation.

* Polignano, M., Basile, P., De Gemmis, M., Semeraro, G., & Basile, V. (2019). Alberto: Italian BERT language understanding model for NLP challenging tasks based on tweets. In *Proceedings of the 6th Italian Conference on Computational Linguistics*, Bari, Italy, November 13- 15, 2019. CEUR Workshop Proceedings 2481, CEUR-WS.org, 2019.

Less represented Languages?



No Italian Language!

a BigScience initiative

BLOOM

176B params · 59 languages · Open-access

* Basile, P., Cassotti, P., Polignano, M., Siciliani, L., & Semeraro, G. (2023). On the impact of Language Adaptation for Large Language Models: A case study for the Italian language using only open resources. In *Proceedings of the 9th Italian Conference on Computational Linguistics*, Venice, Italy, November 30 - December 2, 2023. CEUR Workshop Proceedings 3596, CEUR-WS.org, 2023.

Fail on Italian Culture?

how do I make the perfect carbonara?

To make the perfect carbonara, you will need the following ingredients:

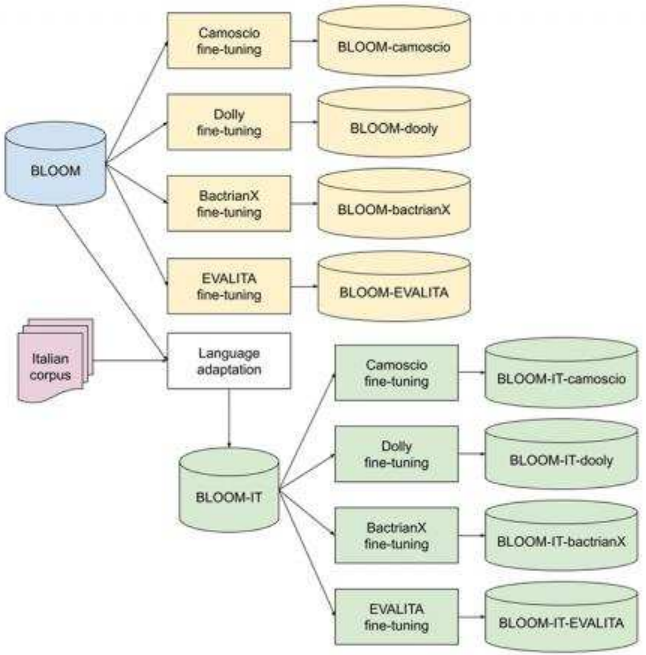
- 8 ounces spaghetti or other pasta
- 4 large eggs
- 1 cup grated Parmesan cheese
- 1/2 cup pancetta or bacon, diced
- 1/2 cup heavy cream
- 2 cloves garlic, minced
- 1/4 teaspoon black pepper
- 1/4 teaspoon salt
- Chopped fresh parsley for garnish (optional)

Here are the steps to make the perfect carbonara:

1. Cook the pasta according to the package instructions until it is al dente (tender but firm

https://x.com/fulviuus/status/1598460506846658560

BLOOM - PEFT + LoRA



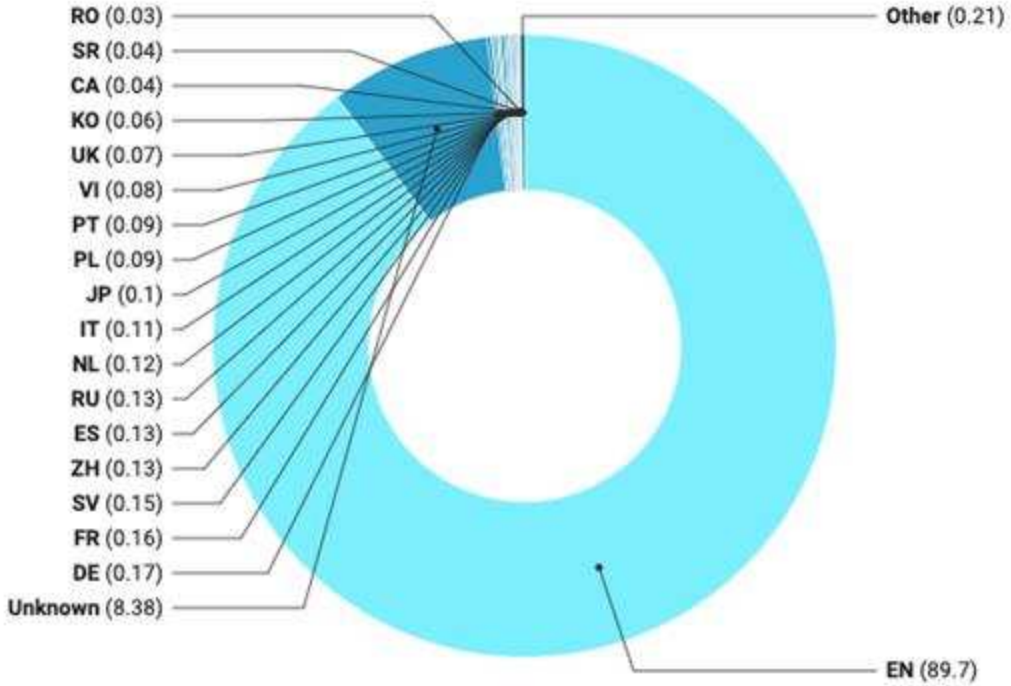
In detail, starting from the BLOOM-1b7 model, we obtain four **fine-tuned models**: one for each instruction dataset (Camoscio, Dolly, and BactrianX) plus the EVALITA model.

Then, the BLOOM-1b7 model is adapted to Italian, leveraging data from the Italian corpus (Italian Wikipedia, Wikinews, and Wikibooks) and obtaining the **Italian-adapted model called BLOOM-IT-1b7**.

BLOOM

* Basile, P., Siciliani, L., Musacchio, E., Polignano, M., & Semeraro, G. (2024). Adapting BLOOM to a new language: A case study for the Italian. *IJCoL. Italian Journal of Computational Linguistics*, *10*(10, 1).

Meta LLaMA 2 same problems as before



90% English pre-training data

Other languages (*German, French, Chinese, Spanish, Dutch, Italian, Japanese, Polish, Portuguese, …*)
less than 2% training data

8% training data “unknown”
(*includes programming code data*)

https://slator.com/meta-warns-large-language-model-may-not-be-suitable-non-english-use/

LLaMAntino

a family of large language models for Italian and its applications

PIERPAOLO BASILE, Università degli Studi di Bari Aldo Moro
ELIO MUSACCHIO, Università degli Studi di Bari Aldo Moro
MARCO POLIGNANO, Università degli Studi di Bari Aldo Moro
LUCIA SICILIANI, Università degli Studi di Bari Aldo Moro
GIUSEPPE FIAMENI, AI & HPC at NVIDIA AI Techn. Center
GIOVANNI SEMERARO, Università degli Studi di Bari Aldo Moro

Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

FAIR FUTURE AI RESEARCH

LLaMAntino

- **Techniques**
  - **Quantization** (*4-bit*)
  - **QLoRA** (*Low-Rank Adaptation*)
  - **FSDP** (*Fully Sharded Data Parallel*)
  - **Argos Translate**: open source offline translation library based on OpenMT
- **Datasets**
  - **Language Adaptation**
    - gsarti/clean_mc4_it medium split
  - **Instruction-Tuning**
    - basilepp19/dolly-15k-it
    - EVALITA 2023 tasks
  - **Chat Fine-Tuning**
    - UltraChat (*translated to Italian*)

Thanks to…

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

NVIDIA

CINECA

consorzio interuniversitario

LEONARDO CINECA

Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

FAIR FUTURE AI RESEARCH

LLaMAntino

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

- *LLaMAntino* is a *family of Italian adapted LLaMA models*
- The family consists of 10 different models, **4** of which are **Italian adapted versions of META - LLaMA base models**:
  - swap-uniba/LLaMAntino-2-7b-hf-ITA
  - swap-uniba/LLaMAntino-2-13b-hf-ITA
  - swap-uniba/LLaMAntino-2-chat-7b-hf-ITA
  - swap-uniba/LLaMAntino-2-chat-13b-hf-ITA
  - swap-uniba/LLaMAntino-2-70b-hf-UltraChat-ITA

- **Goal:** Provide Italian researchers with LLMs that show a *good understanding of the Italian language*
- Should be **further tuned** to improve their capabilities on **specific tasks** …

LLaMAntino whitepaper: https://arxiv.org/abs/2312.09993

LLaMAntino

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

All models were trained on the **Leonardo HPC**

Language Adaptation	Fine-tuning
4-bit quantization, QLoRA, SFTTrainer	Fully-Sharded Data Parallel (FSDP)
3 nodes for a total of **12 GPUs A100 64GB**	**2 nodes** for a total of **8 GPUs A100 64GB**
LoRA parameters: attention dimension (64), scaling parameter (16), dropout (0.1). Single GPU batch size (8). Steps (25K) Text length of (1024)	Single GPU batch size (16). Epochs (3 for 7B, 5 for 13B). Text length (1024)
~100.000 Leonardo hours	**~50.000 Leonardo hours**

LLaMAntino 

Chat Models

LIMITS

- **Hardware:** 8/12 Nvidia A100 GPUs - 512GB PC RAM
- **Data Amount:** 150-500k dialogues or Q/A in native language
- **Grammatical Errors Propagation** if Automatic Translator used for data
- Answers provided for topics outside specific task scope
- Biases in answers

Hallucinations …

LLaMAntino - ANita

SFT on LLaMA-3

https://huggingface.co/swap-uniba/LLaMAntino-3-ANITA-8B-Inst-DPO-ITA

* https://arxiv.org/pdf/2405.07101

Stanford Alpaca

Databricks' Dolly 2.0

WizardLM

LLaMAntino - ANita

DPO on LLaMAntino

https://huggingface.co/swap-uniba/LLaMAntino-3-ANITA-8B-Inst-DPO-ITA
* https://arxiv.org/pdf/2405.07101

mlabonne/orpo-dpo-mix-40k

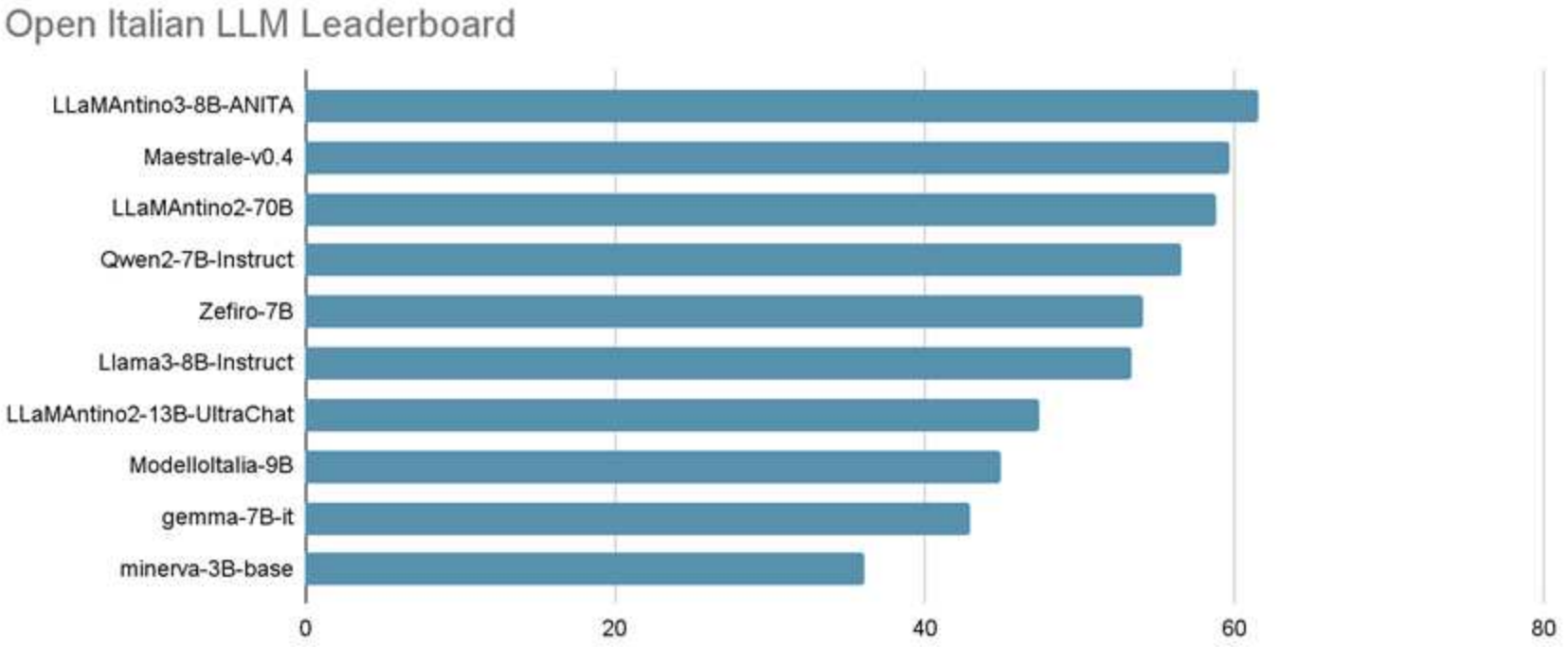
LLaMAntino-3-ANITA-8B-Inst-DPO-ITA is a model of the LLaMAntino - *Large Language Models family*. The model is an instruction-tuned version of Meta-Llama-3-8b-instruct (a fine-tuned **LLaMA 3 model**). This model version aims to be the a Multilingual Model 🏁 (EN 🇺🇸 + ITA🇮🇹) to further fine-tuning on Specific Tasks in Italian.

SCAN ME

https://chat.llamantino.it/

LLaMAntino

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO



https://huggingface.co/spaces/FinancialSupport/open_ita_llm_leaderboard

Large Agentic Models (LAMs)

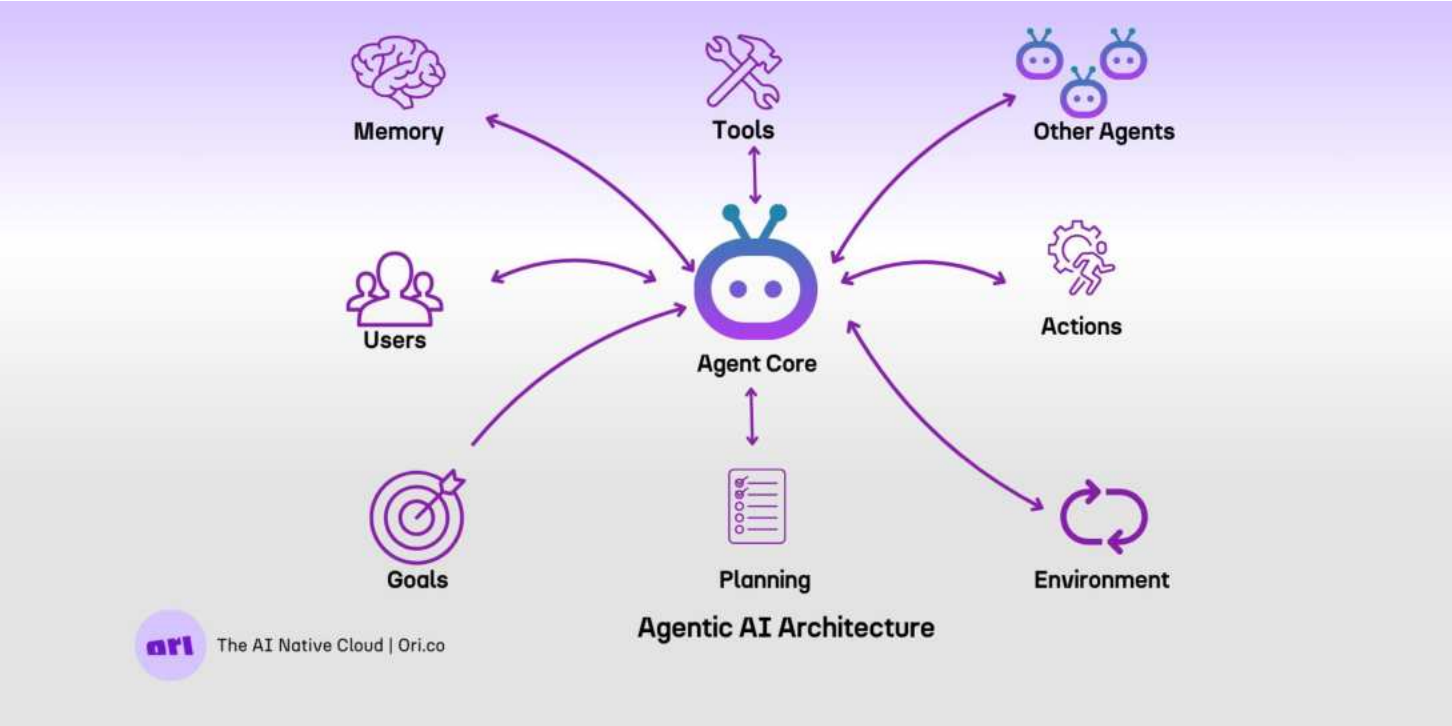
Agentic AI

«Nearly half (48%) of all consumers say they would interact with AI more frequently if it would enhance their experience. This is where **Agentic AI shines**, making generative AI more actionable, contextual, and autonomous.» *https://blog.ori.co/ai-agent-introduction

An autonomous agent is a system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and to affect what it senses in the future.

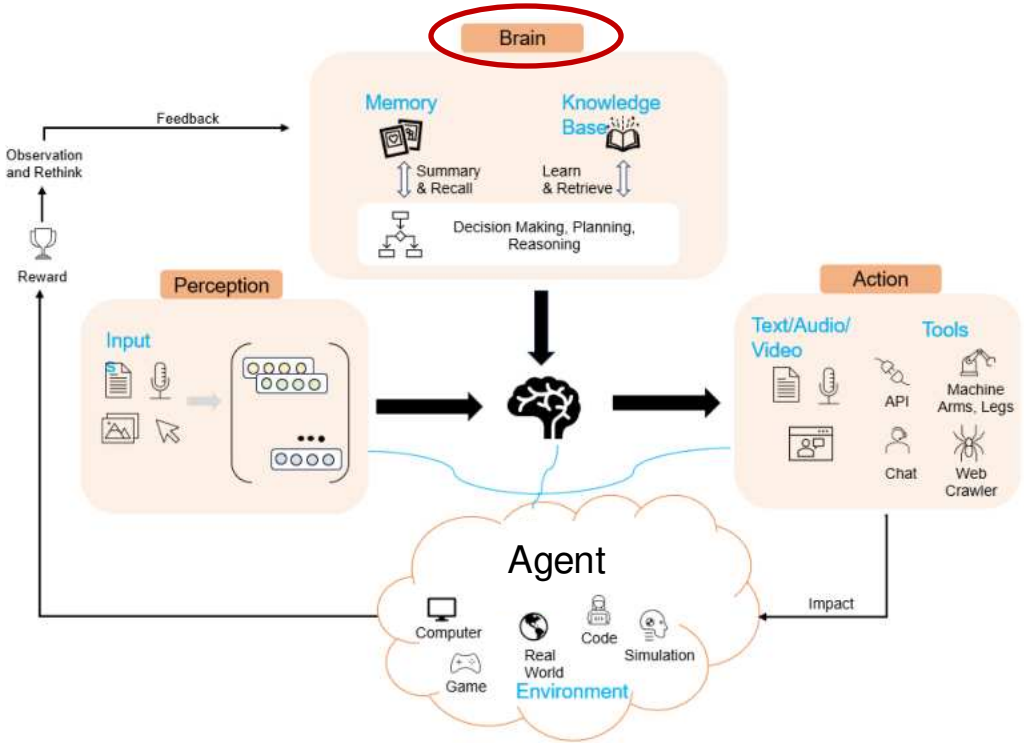
Franklin, S. (1997). Autonomous agents as embodied AI. Cybernetics & Systems, 28(6), 499-520.

Agentic AI



https://blog.ori.co/ai-agent-introduction

What's next? Large Agentic Models (LAMs)



Rupali Patil: https://medium.com/towards-artificial-intelligence/whats-emerging-in-ai-autonomous-multi-agents-and-large-action-agentic-models-lams-7e882a659565

Large Agentic Models (LAMs)

They encompass models that exhibit **agency** to **act independently** within their environment. Such models learn from interaction with the real world, perform capabilities like planning and decision-making, and take action. The large agentic model provides a framework for autonomous agents to interact with each other and the environment and to adapt their behavior based on feedback and learning.

Examples: *Agents that can plan, make decisions, use tools and act autonomously* ***interacting with the environment and even other agents or AI models…***

Rupali Patil: https://medium.com/towards-artificial-intelligence/whats-emerging-in-ai-autonomous-multi-agents-and-large-action-agentic-models-lams-7e882a659565

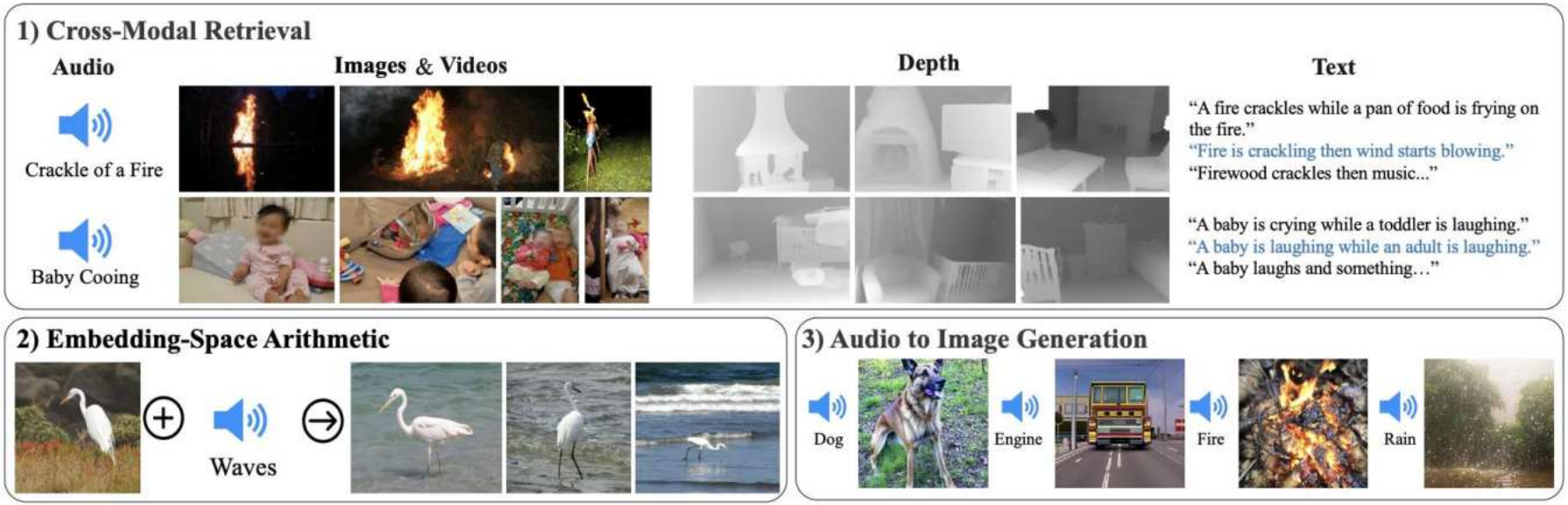
LAMs

Rupali Patil: https://medium.com/towards-artificial-intelligence/whats-emerging-in-ai-autonomous-multi-agents-and-large-action-agentic-models-lams-7e882a659565

	Large Language Model (LLM)	Large Action Model (LAM)	Large Agentic Model (LAM)
Focus	Language understanding and generation	Taking actions in the real world	Acting independently with agency
Core capability	Process and generate text, answer questions	Perform physical actions, manipulate objects, or influence the real world	Complex reasoning, decision-making, and take actions
Data used in training	Massive amounts of text and code	May include sensor data, user interaction data, and real-world observations	May include sensor data, user interaction data, and real-world observations
Learns from	Pattern recognition from large datasets used in training	Observation and demonstration of user's actions	Pattern recognition, self-assessment, and interaction with real world
Reasoning ability	Limited to single-step reasoning based on language patterns and knowledge base	Zero-shot reasoning capabilities of a neural network without any prior training on specific task	Advanced multi-step reasoning based on context and problem breakdown
Use Cases	Content creation, Q/A, language translations, chatbots	Task automation, personalized assistance, streamlining workflows, enhanced customer service	Industrial automation, supply chain management and logistics, financial trading, personalized investments

Perception for LAMs

LAMs can become more effective agents by perceiving the environment.

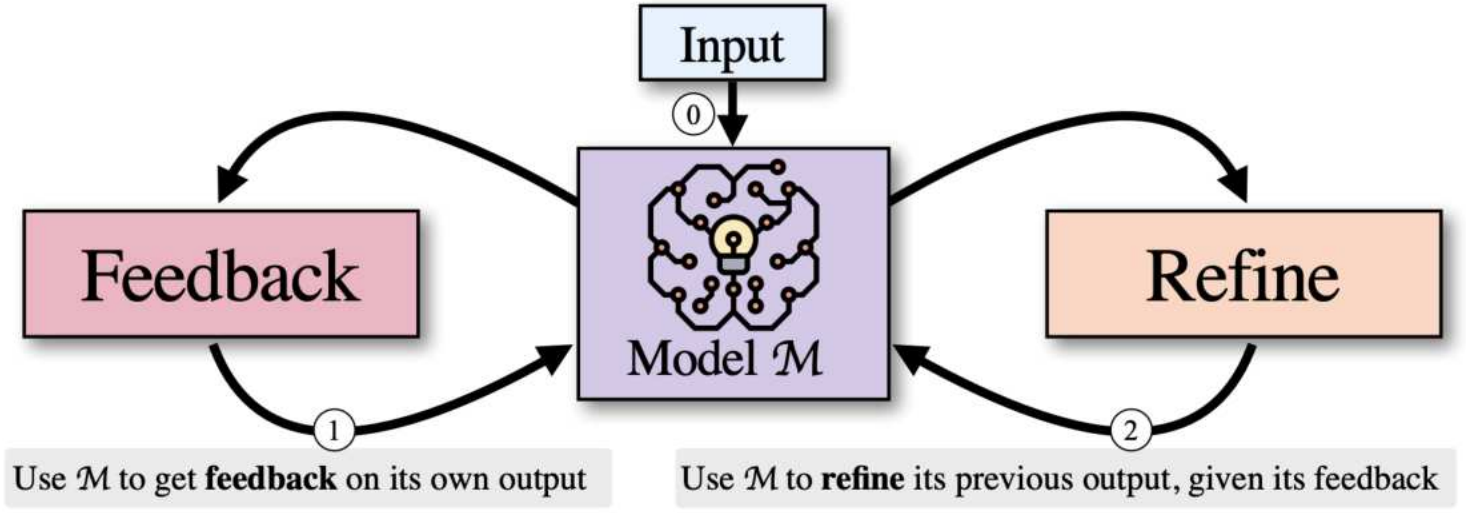


Girdhar, Rohit, et al. "Imagebind: One embedding space to bind them all." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

Self-Refine for LAMs

LAMs can become more effective agents by reflecting on their own behavior.

Like humans, large language models (LLMs) do not always generate the best output on their first try.



"Self-Refine: Iterative Refinement with Self-Feedback," Madaan et al. (2023)

Tool Evaluation for LAMs

LAMs can become more effective agents by evaluating their own behavior.

LLMs should interact with appropriate tools to evaluate certain aspects of the text and then revise the output based on the feedback obtained during this validation process.

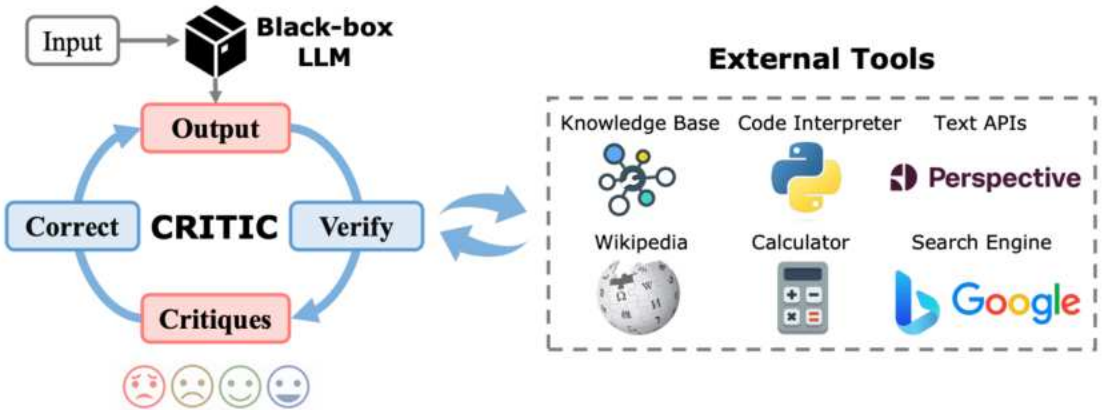


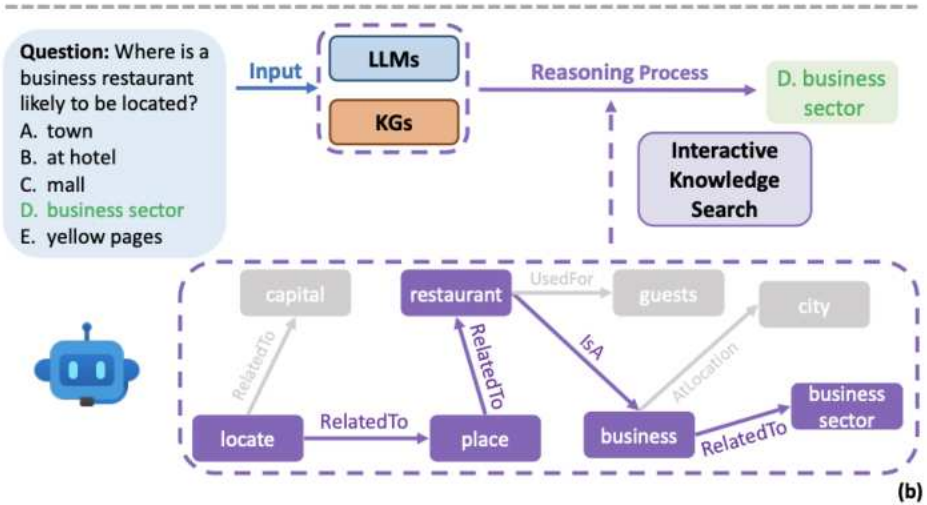
Figure 1: The CRITIC framework consists of two steps: (1) verifying the output by interacting with external tools to generate critiques and (2) correcting the output based on the received critiques. We can iterative such *verify-then-correct* process to enable continuous improvements.

"CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing," Gou et al. (2024)

Access to Knowledge in LAMs (Neuro-Symbolic)

LAMs can become more effective agents by accessing verified information

LLMs should interact with knowledge bases to extract relevant information already codified and verified by humans. If structured over graphs or symbolic rules inference can be done.



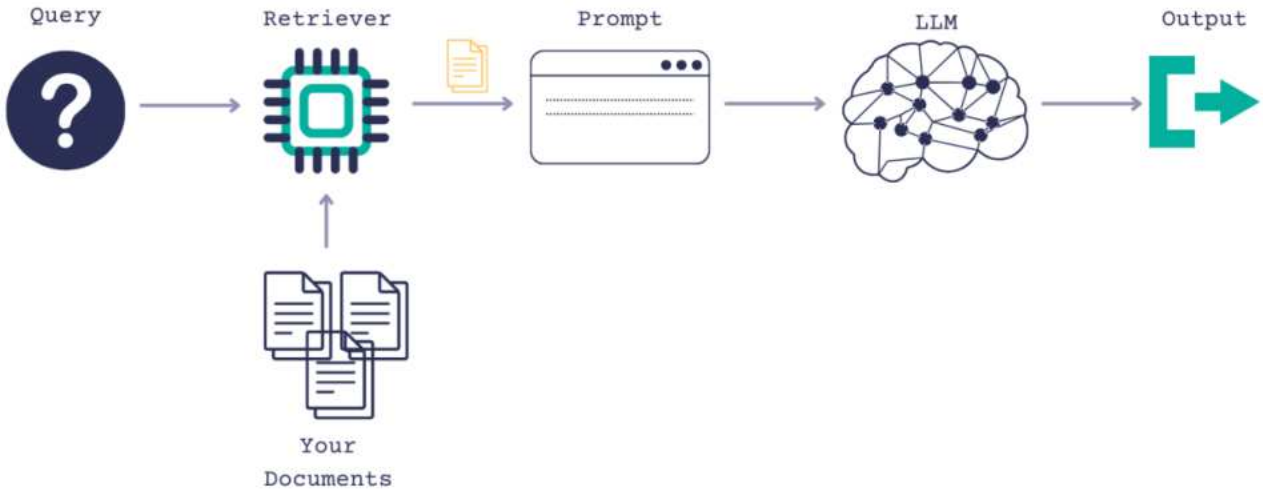
Feng, C., Zhang, X., & Fei, Z. (2023). Knowledge solver: Teaching llms to search for domain knowledge from knowledge graphs. *arXiv preprint arXiv:2309.03118*.

Calanzone, Diego, Stefano Teso, and Antonio Vergari. "Logically Consistent Language Models via Neuro-Symbolic Integration." arXiv preprint arXiv:2409.13724 (2024).

Access to Knowledge in LAMs (RAG)

LAMs can become more effective agents by accessing verified information

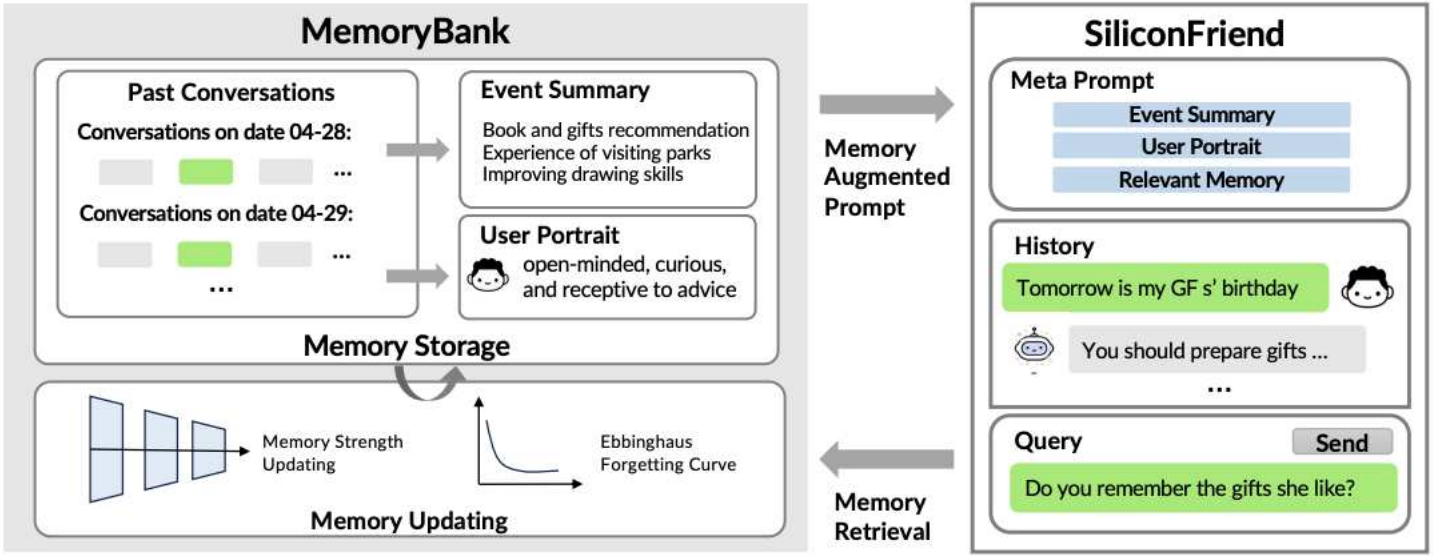
Tool Use, in which an LLM is given functions it can request to call for gathering timely information, taking action, or manipulating data, is a key design pattern of AI agentic workflows.



Gao, Yunfan, et al. "Retrieval-augmented generation for large language models: A survey." *arXiv preprint arXiv:2312.10997* (2023).

Access to Memory in LAMs

LAMs can become more effective agents by accessing past experience



Zhong, Wanjun, et al. "Memorybank: Enhancing large language models with long-term memory." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. No. 17. 2024.

Personalization in LAMs

LAMs can become more effective agents by personalizing the actions

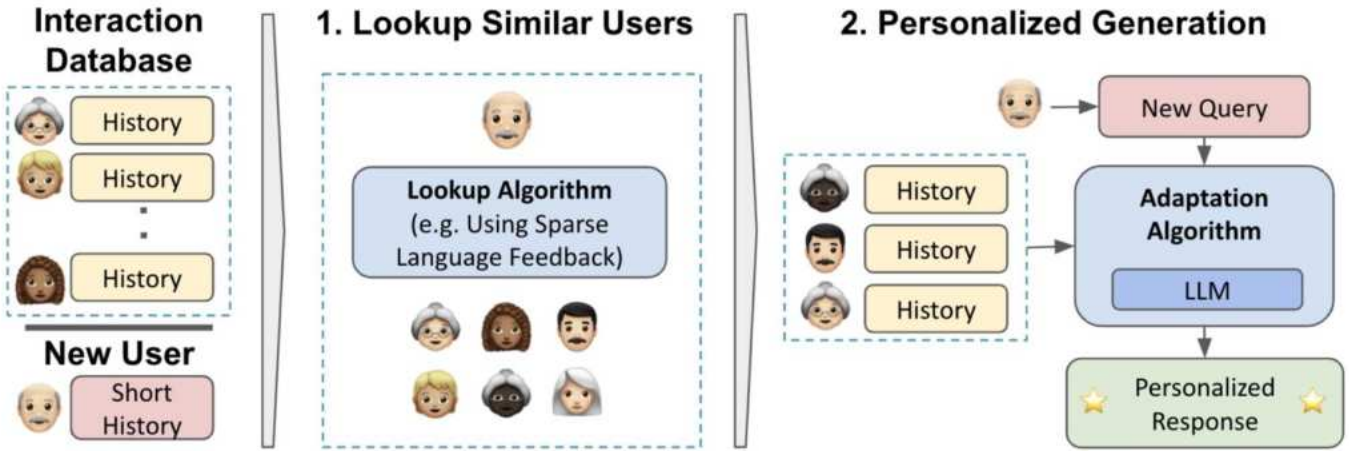


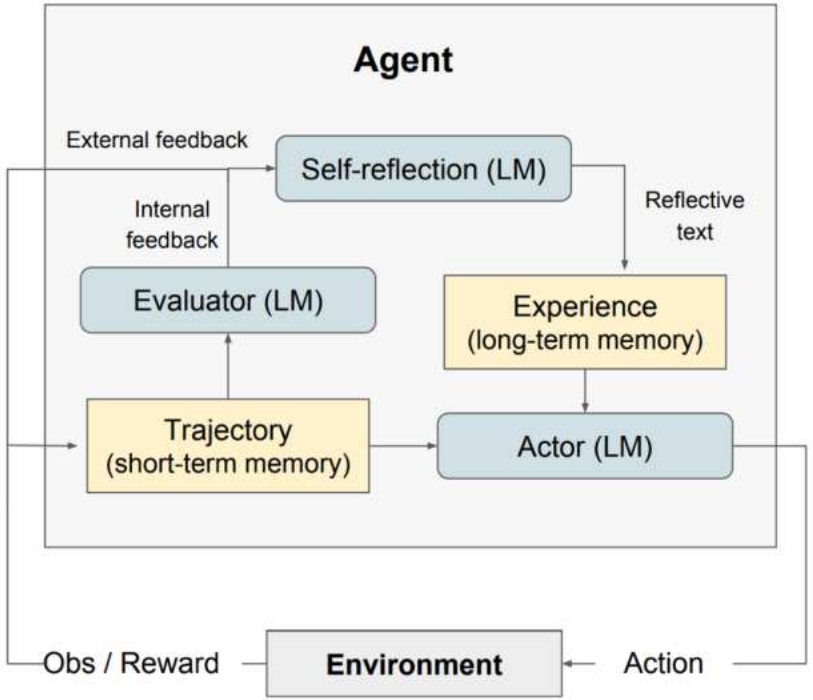
Figure 2: In the canonical personalization setting, a dataset of historical users and their interactions is leveraged to personalize interactions for a new user with a limited history. PersonalLLM enables the development of such methods for learning *across* users.

Zollo, Thomas P., et al. "PersonalLLM: Tailoring LLMs to Individual Preferences." *arXiv preprint arXiv:2409.20296* (2024).

Reflexion for LAMs

LAMs can become more effective agents by short-long term goals

It remains challenging for these language agents to quickly and efficiently learn from trial-and-error as traditional reinforcement learning methods require extensive training samples and expensive model fine-tuning.



"Reflexion: Language Agents with Verbal Reinforcement Learning," Shinn et al. (2023)

Planning in LAMs

LAMs can become more effective agents by planning the steps to perform

Planning is a key agentic AI design pattern in which we use a large language model (LLM) to autonomously decide on what sequence of steps to execute to accomplish a larger task.

Many tasks can't be done in a single step or with a single tool invocation, but an agent can decide what steps to take.

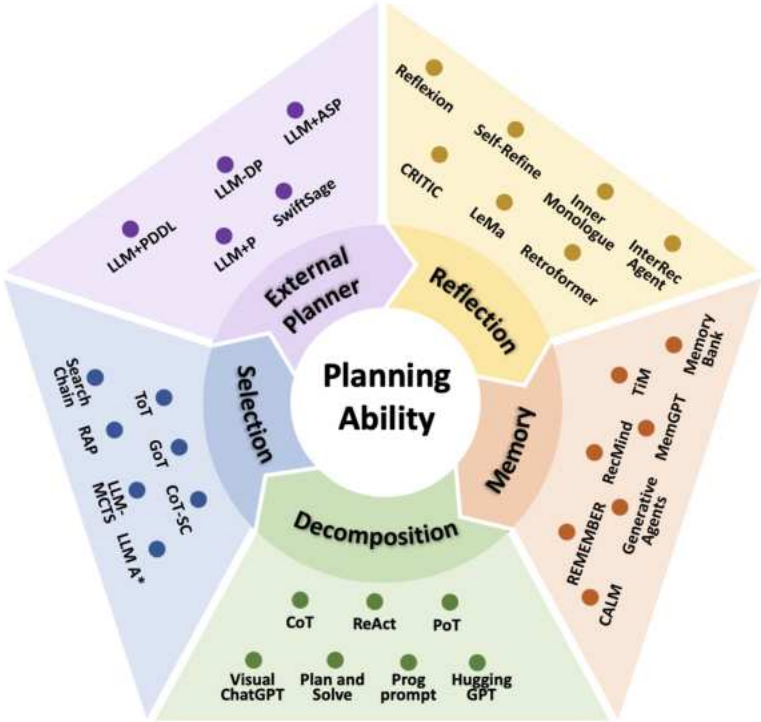


Figure 1: Taxonomy on LLM-Agent planning.

"Understanding the planning of LLM agents: A survey," by Huang et al. (2024)

Acting as a LAMs

LAMs can become more effective agents by interact with available systems

Large Language Models (LLMs) have seen an impressive wave of advances recently, with models now excelling in a variety of tasks, such as mathematical reasoning and program synthesis. However, their potential to effectively use tools via API calls remains unfulfilled.

"Gorilla: Large Language Model Connected with Massive APIs," Patil et al. (2023)

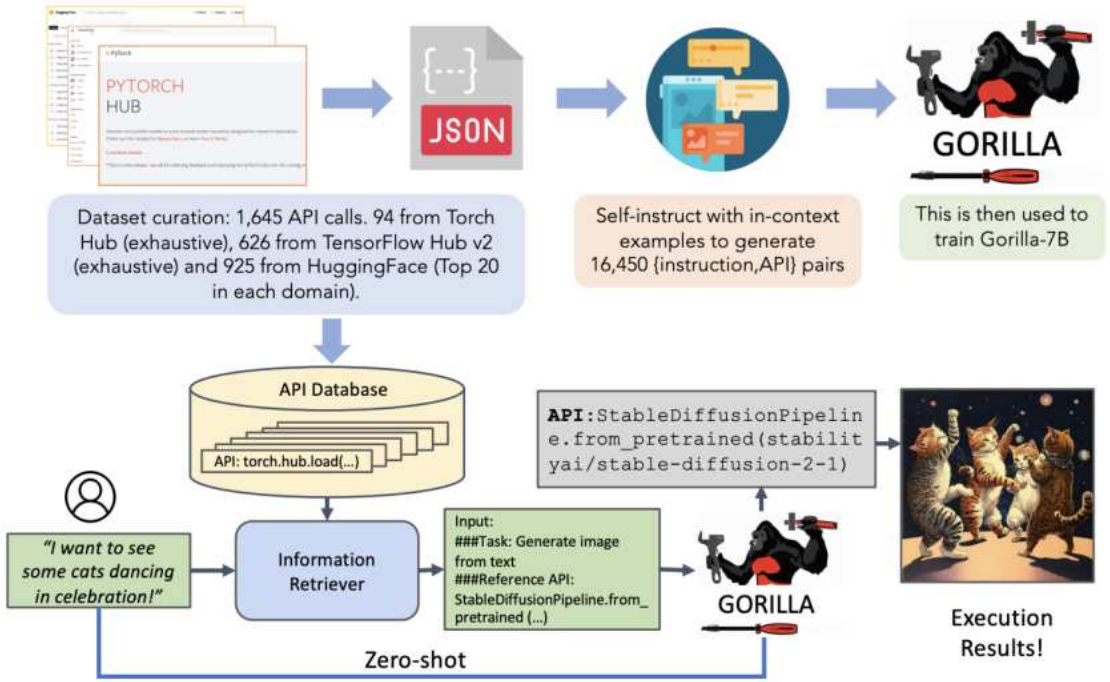
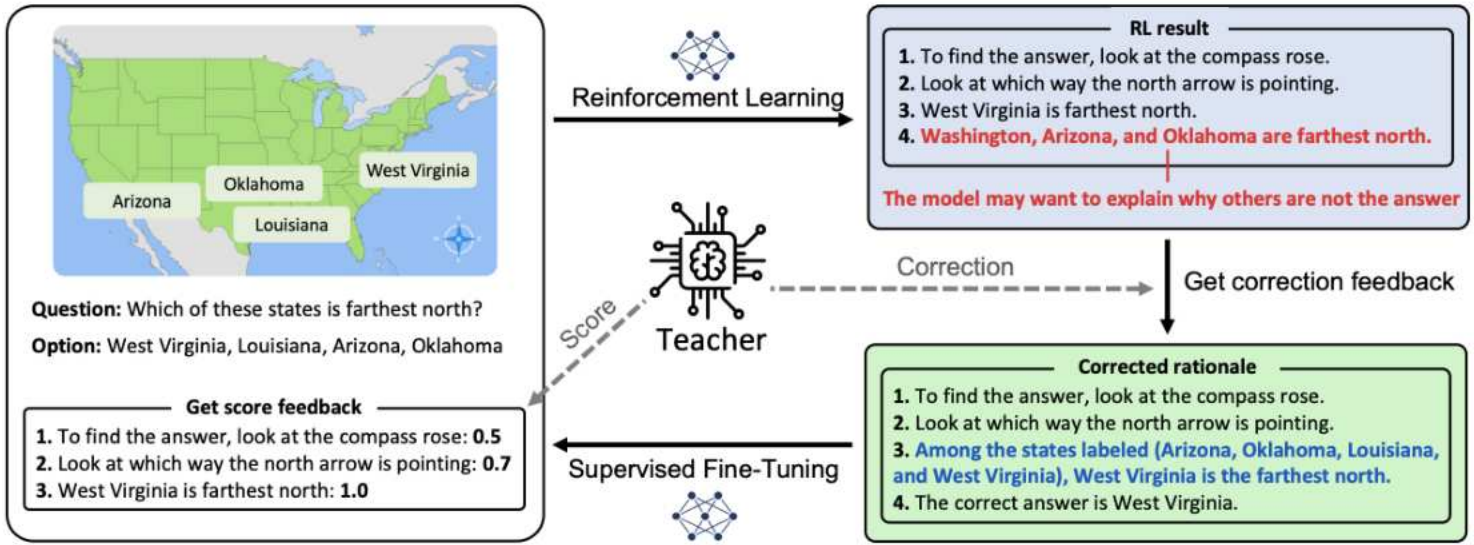


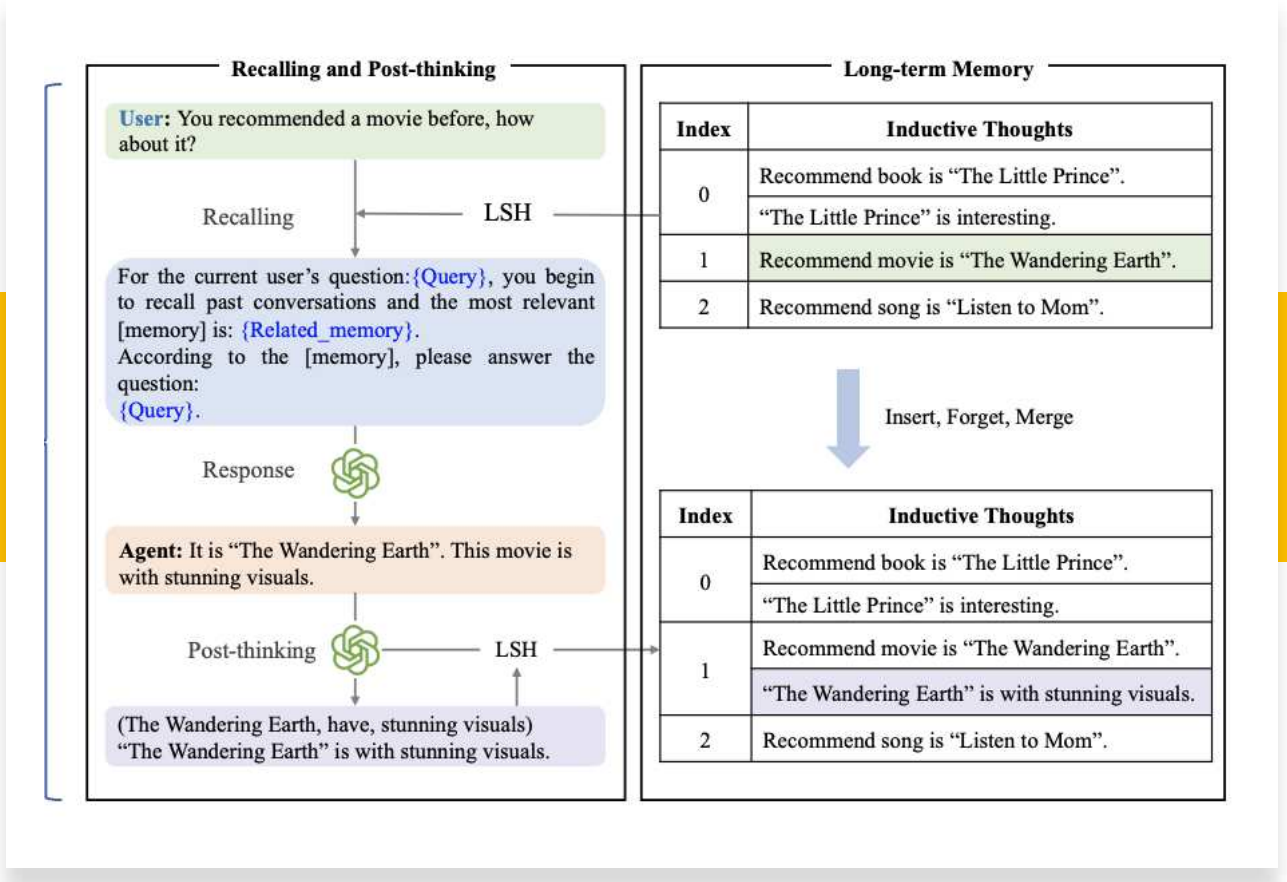
Figure 3: **Gorilla: A system for enabling LLMs to interact with APIs.** The upper half represents the training procedure as described in Sec 3. This is the most exhaustive API data-set for ML to the best of our knowledge. During inference (lower half), Gorilla supports two modes - with retrieval, and zero-shot. In this example, it is able to suggest the right API call for generating the image from the user's natural language query.

Learn from feedback in LAMs (finetuning)

LAMs can become more effective agents by learning from feedback



Byun, Ju-Seung, et al. "ARES: Alternating Reinforcement Learning and Supervised Fine-Tuning for Enhanced Multi-Modal Chain-of-Thought Reasoning Through Diverse AI Feedback." *arXiv preprint arXiv:2407.00087* (2024).



Learn from feedback in LAMs (memory update)

LAMs can become more effective agents by learning from past

Liu, Lei, et al. "Think-in-memory: Recalling and post-thinking enable llms with long-term memory." *arXiv preprint arXiv:2311.08719* (2023).

Why we should move to M-LAMs?

Some tasks are easier to improve via system design and dedicated systems.

While LLMs appear to follow remarkable scaling laws that predictably yield better results with more compute, in many applications, scaling offers lower returns-vs-cost than building a compound system.

FROM THE MAKERS OF WOLFRAM LANGUAGE AND MATHEMATICA

* Zaharia, Matei, et al. "The shift from models to compound ai systems." *Berkeley Artificial Intelligence Research Lab. Available online at: https://bair. berkeley. edu/blog/2024/02/18/compound-ai-systems/(accessed February 27, 2024)* (2024).

Why we should move to M-LAMs

Performance goals vary widely.

	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4	GPT-3.5	Gemini 1.0 Ultra	Gemini 1.0 Pro
Undergraduate level knowledge *MMLU*	86.8% 5 shot	79.0% 5-shot	75.2% 5-shot	86.4% 5-shot	70.0% 5-shot	83.7% 5-shot	71.8% 5-shot
Graduate level reasoning *GPQA, Diamond*	50.4% 0-shot CoT	40.4% 0-shot CoT	33.3% 0-shot CoT	35.7% 0-shot CoT	28.1% 0-shot CoT	–	–
Grade school math *GSM8K*	95.0% 0-shot CoT	92.3% 0-shot CoT	88.9% 0-shot CoT	92.0% 5-shot CoT	57.1% 5-shot	94.4% Maj1@32	86.5% Maj1@32
Math problem-solving *MATH*	60.1% 0-shot CoT	43.1% 0-shot CoT	38.9% 0-shot CoT	52.9% 4-shot	34.1% 4-shot	53.2% 4-shot	32.6% 4-shot
Multilingual math *MGSM*	90.7% 0-shot	83.5% 0-shot	75.1% 0-shot	74.5% 8-shot	–	79.0% 8-shot	63.5% 8-shot
Code *HumanEval*	84.9% 0-shot	73.0% 0-shot	75.9% 0-shot	67.0% 0-shot	48.1% 0-shot	74.4% 0-shot	67.7% 0-shot
Reasoning over text *DROP, F1 score*	83.1 3-shot	78.9 3-shot	78.4 3-shot	80.9 3-shot	64.1 3-shot	82.4 Variable shots	74.1 Variable shots
Mixed evaluations *BIG-Bench-Hard*	86.8% 3-shot CoT	82.9% 3-shot CoT	73.7% 3-shot CoT	83.1% 3-shot CoT	66.6% 3-shot CoT	83.6% 3-shot CoT	75.0% 3-shot CoT
Knowledge Q&A *ARC-Challenge*	96.4% 25-shot	93.2% 25-shot	89.2% 25-shot	96.3% 25-shot	85.2% 25-shot	–	–
Common Knowledge *HellaSwag*	95.4% 10-shot	89.0% 10-shot	85.9% 10-shot	95.3% 10-shot	85.5% 10-shot	87.8% 10-shot	84.7% 10-shot

https://www.anthropic.com/news/claude-3-family

* Zaharia, Matei, et al. "The shift from models to compound ai systems." *Berkeley Artificial Intelligence Research Lab. Available online at: https://bair. berkeley. edu/blog/2024/02/18/compound-ai-systems/(accessed February 27, 2024)* (2024).

Why we should move to M-LAMs?

Systems can be dynamic.

LLMs are inherently limited because they are trained on static datasets, so their "knowledge" is fixed.

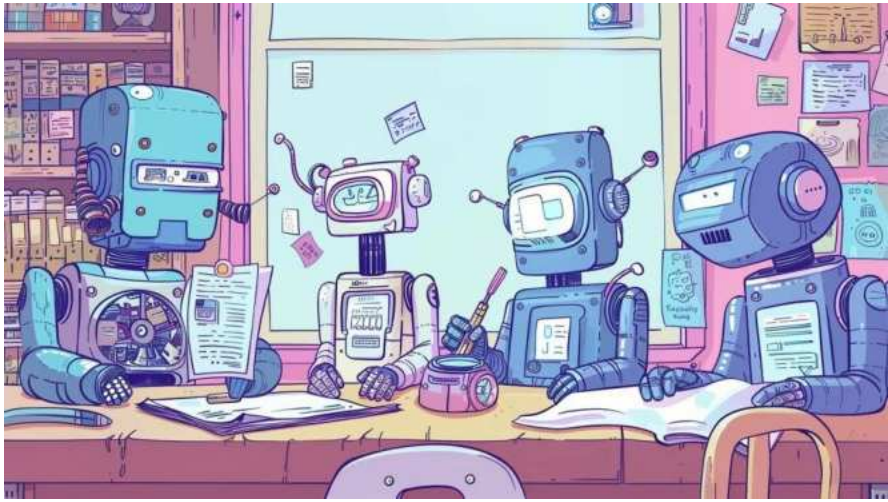
Therefore, developers need to combine models with other components, **such as search and retrieval (RAG),** to incorporate **timely data**.

In addition, training lets a model "see" the whole training set, so more complex systems are needed to build AI applications with access controls (e.g., answer a user's questions based only on files the user has access to).

Why we should move to M-LAMs?

Cooperation among specialized LAMs can improve performances

When multiple AI agents collaborate, the results are often superior. This could be in coding, planning, creative writing, or any other domain. The iterative feedback loop between agents ensures that the final output is refined and of high quality.



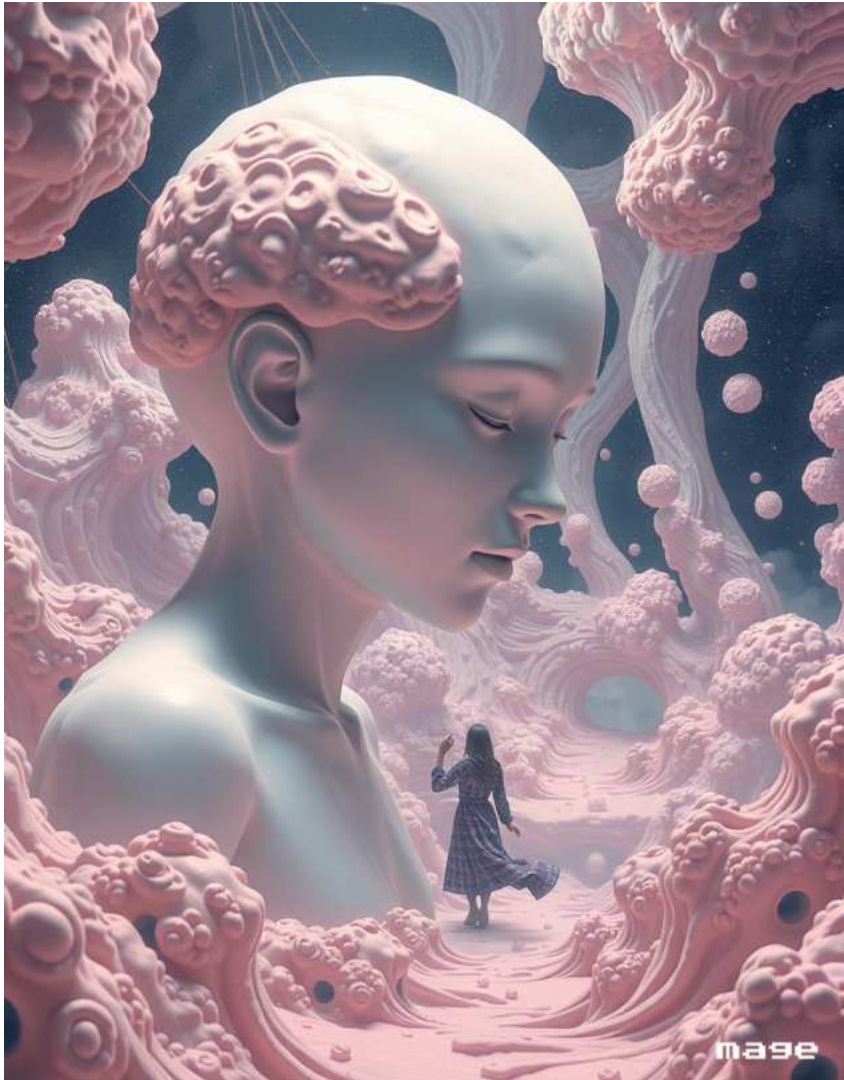
Wu, Qingyun, et al. "Autogen: Enabling next-gen llm applications via multi-agent conversation framework." *arXiv preprint arXiv:2308.08155* (2023).

Why we should move to M-LAMs?

Improving control and trust is easier with systems.

Using an AI system instead of a model can help developers control behavior more tightly, e.g., by filtering model outputs.

Likewise, even the best LLMs still hallucinate, but a system combining, say, LLMs with retrieval can increase user trust by providing citations or automatically verifying facts.



Multi agent systems (MAS)

The Multi-Agent approach to solving problems has been around for quite some time and has some very sound theoretical grounding. The origins of MAS can be traced to the 1970s and 1980s when researchers began to explore the idea of Distributed Artificial Intelligence (DAI).

Ferbes and Weiss, in their book: ***«Multi-Agent Systems: An Introduction to Distributed Artificial Intelligence»*** *provide* a **comprehensive guide on multi-agent systems (MAS)** and distributed artificial intelligence (DAI).

The book explores how independent agents autonomous units with their own goals and actions can work together within a shared environment. The text provides an in-depth look at how MAS can be applied to solve complex problems that a single-agent system might struggle with, such as resource allocation, scheduling, robotics, and more.

*Ferber, J., & Weiss, G. (1999). *Multi-agent systems: an introduction to distributed artificial intelligence* (Vol. 1). Reading: Addison-wesley.

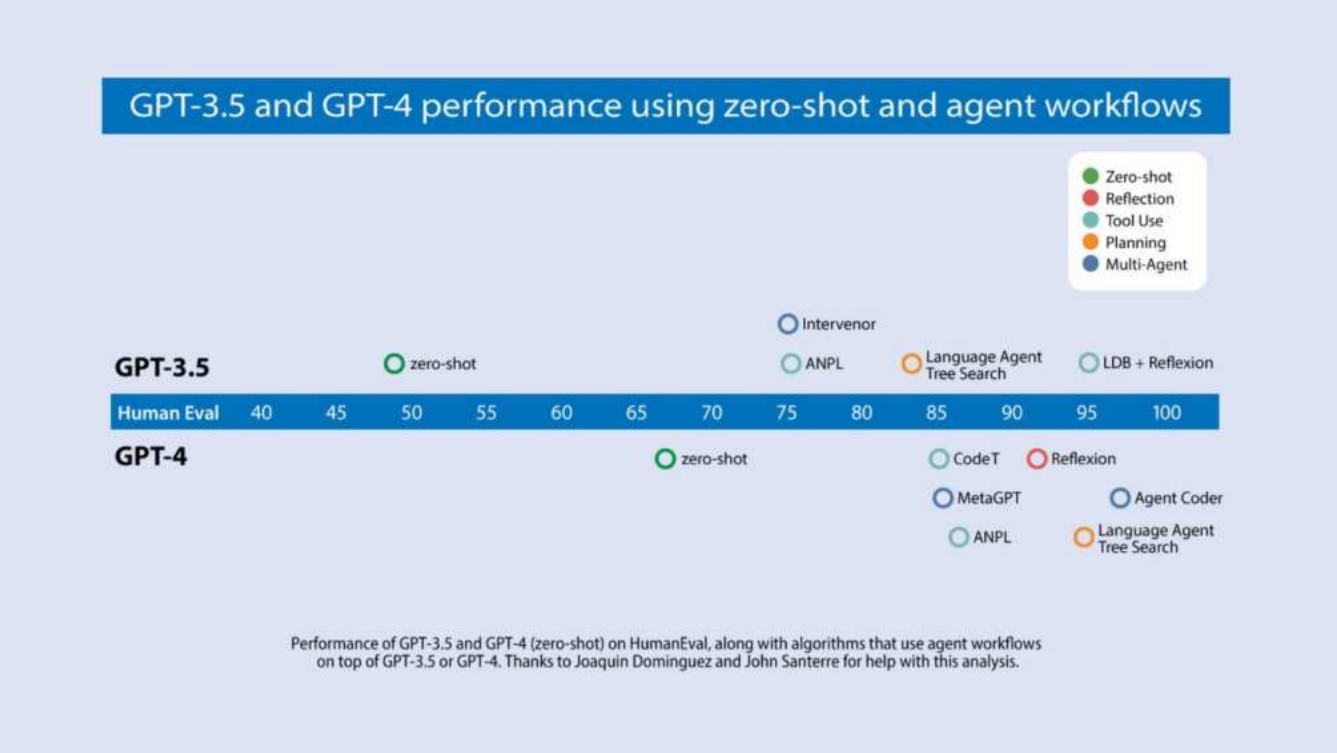
Multi agent systems (MAS)

Table 2.1 Classification of interaction situations.

Goals	Resources	Skills	Types of situation	Category
Compatible	Sufficient	Sufficient	*Independence*	Indifference
Compatible	Sufficient	Insufficient	*Simple collaboration*	
Compatible	Insufficient	Sufficient	*Obstruction*	Cooperation
Compatible	Insufficient	Insufficient	*Coordinated collaboration*	
Incompatible	Sufficient	Sufficient	*Pure individual competition*	
Incompatible	Sufficient	Insufficient	*Pure collective competition*	Antagonism
Incompatible	Insufficient	Sufficient	*Individual conflicts over resources*	
Incompatible	Insufficient	Insufficient	*Collective conflicts over resources*	

*Ferber, J., & Weiss, G. (1999). *Multi-agent systems: an introduction to distributed artificial intelligence* (Vol. 1). Reading: Addison-wesley.

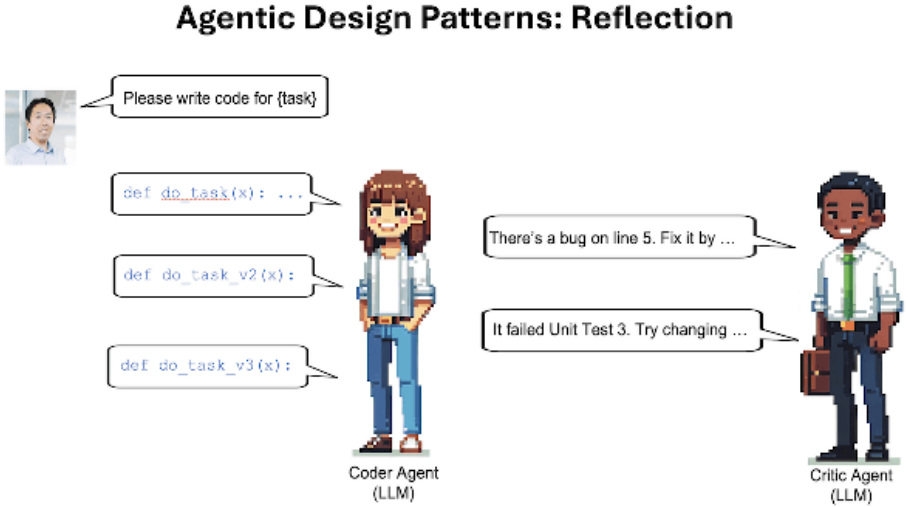
MAS – the Andrew Ng overview (2024)



https://www.deeplearning.ai/the-batch/how-agents-can-improve-llm-performance/

https://www.deeplearning.ai/the-batch/agentic-design-patterns-part-2-reflection/

Judge in MAS



Intervenor

INTERVENOR prompts Large Language Models (LLMs) to play distinct roles during the code repair process, functioning as both a Code Learner and a Code Teacher.

Teacher is responsible for crafting a **Chain-of-Repair** (CoR) to serve as guidance for the Code Learner. During generating the CoR, the Code Teacher needs to check the generated codes from Code Learner and reassess how to address code bugs based on error feedback received from compilers

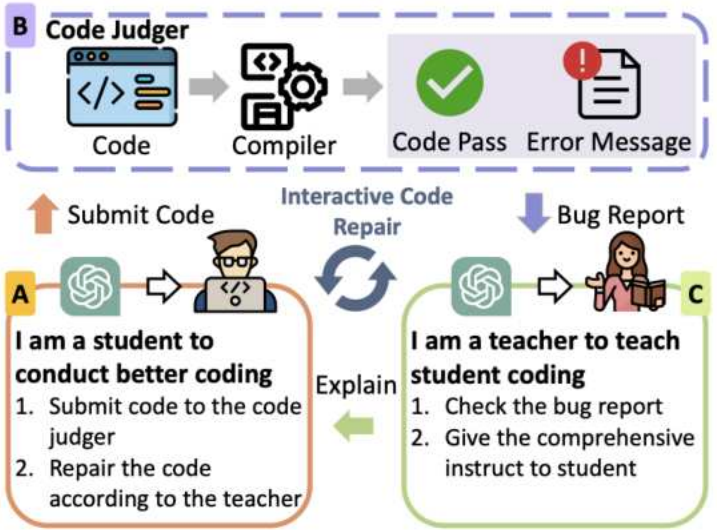


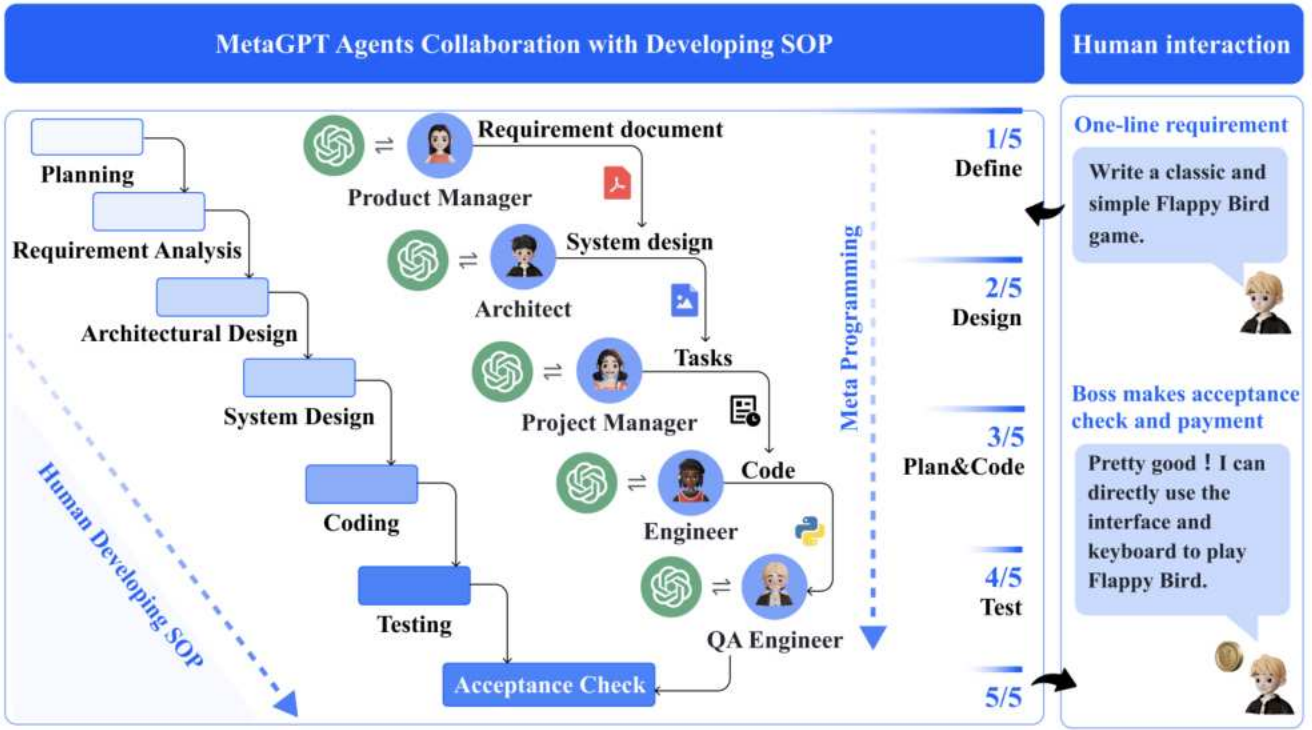
Figure 1: The Illustration of INTERVENOR. There are two agents in INTERVENOR, the teacher and student, who collaborate to repair the code. The error messages are utilized as a kind of INTERVENOR to alleviate the Degeneration-of-Thought (DoT) problem.

Wang, Hanbin, et al. "INTERVENOR: Prompting the Coding Ability of Large Language Models with the Interactive Chain of Repair." *Findings of the Association for Computational Linguistics ACL 2024*. 2024.

MetaGPT

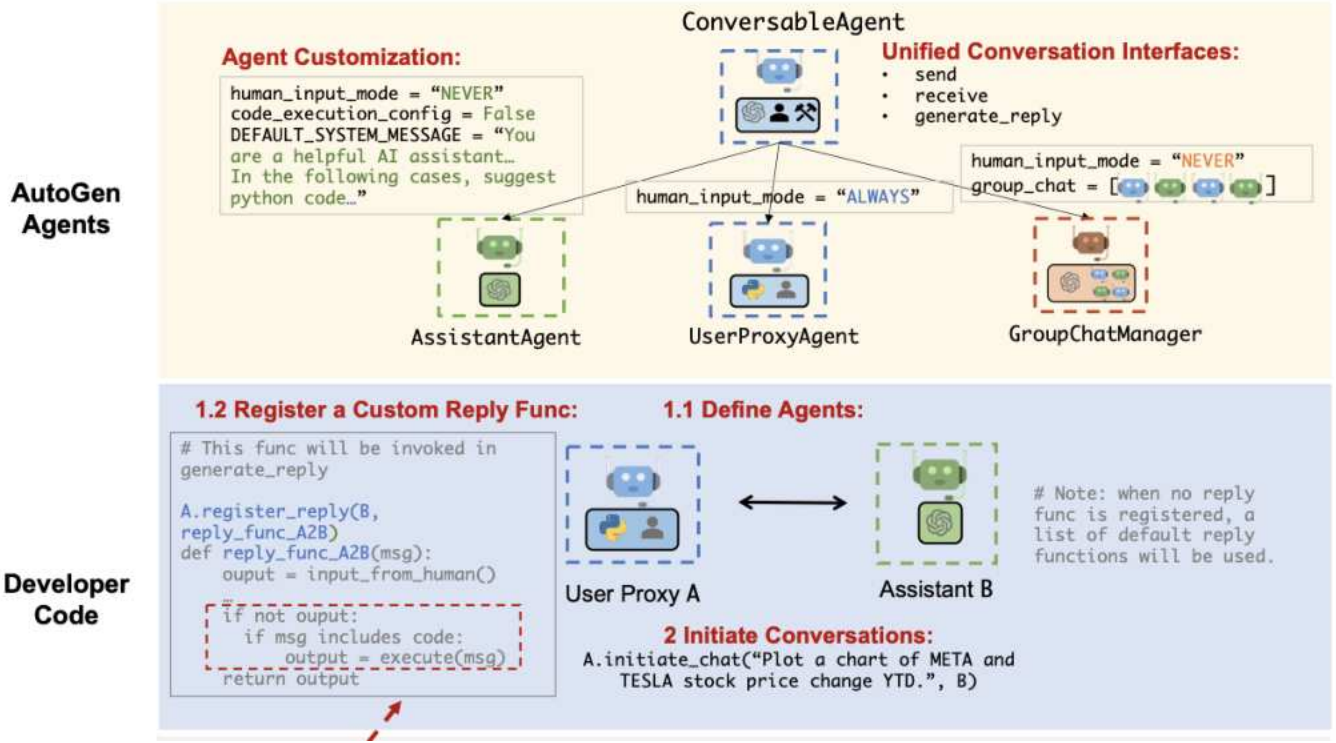
The software development SOPs between MetaGPT and real-world human teams. In software engineering, SOPs promote collaboration among various roles.

MetaGPT showcases its ability to decompose complex tasks into specific actionable procedures assigned to various roles (e.g., Product Manager, Architect, Engineer, etc.).



Hong, Sirui, et al. "Metagpt: Meta programming for multi-agent collaborative framework." *arXiv preprint arXiv:2308.00352* (2023).

AutoGen: Enabling Next-Gen LLM

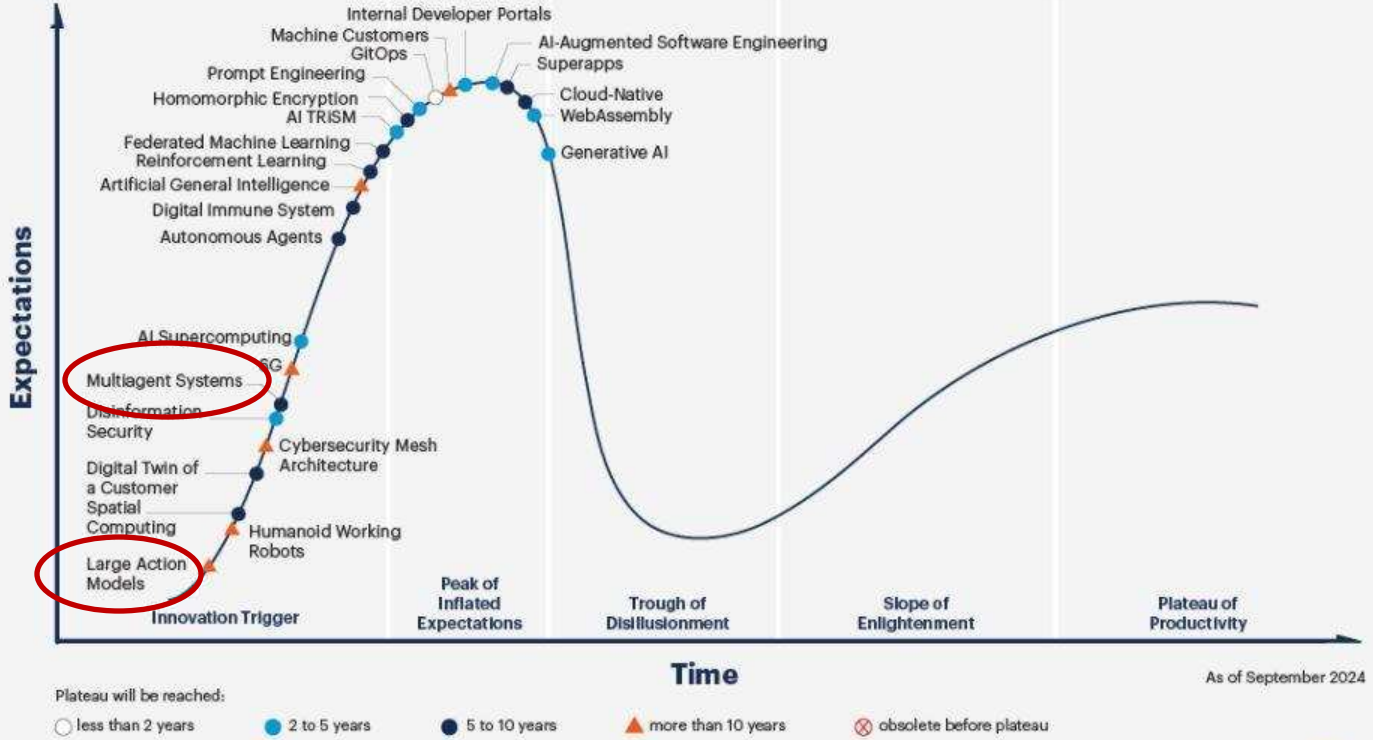


Wu, Qingyun, et al. "Autogen: Enabling next-gen llm applications via multi-agent conversation framework." *arXiv preprint arXiv:2308.08155* (2023).

IS THE FUTURE AGENTIC?

Marco Polignano, Marco de Gemmis and Giovanni Semeraro. Unraveling the Enigma of SPLIT in Large-Language Models: The Unforeseen Impact of System Prompts on LLMs with Dissociative Identity Disorder. In proceedings of Tenth Italian Conference on Computational Linguistics, Pisa, 4 - 6 December 2024

Hype Cycle for Emerging Technologies, 2024



Source: Gartner

Gartner

Thank You!

Marco Polignano – RTD-A FAIR
Dip. di Informatica
Università degli Studi di Bari Aldo Moro

marco.polignano@uniba.it

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO