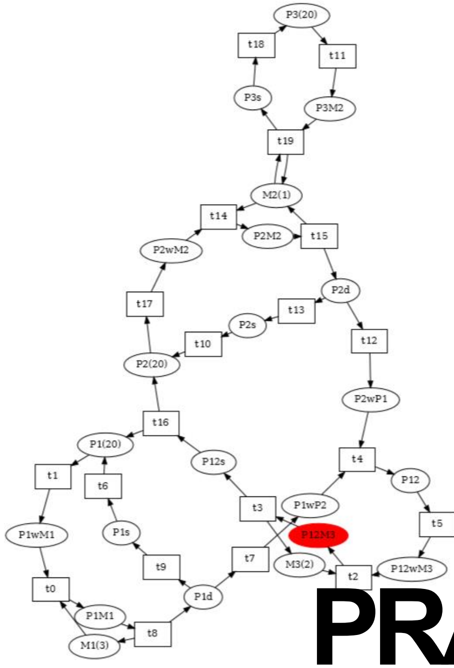
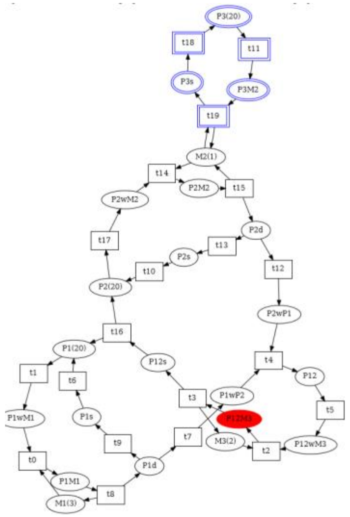
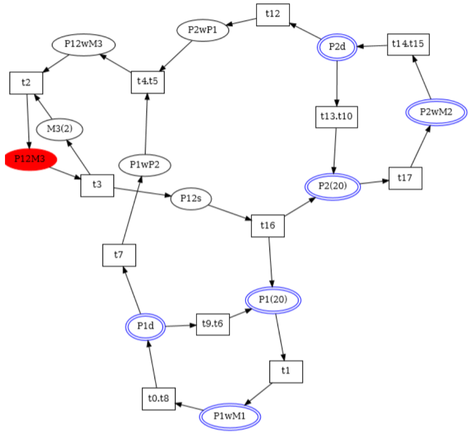
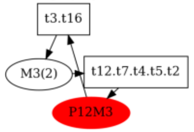








PRACTICAL LESSONS FROM APPLYING LARGE LANGUAGE MODELS IN TEACHING, DEVELOPMENT, AND RESEARCH

Yann Thierry-Mieg

LIP6, Sorbonne Université, CNRS

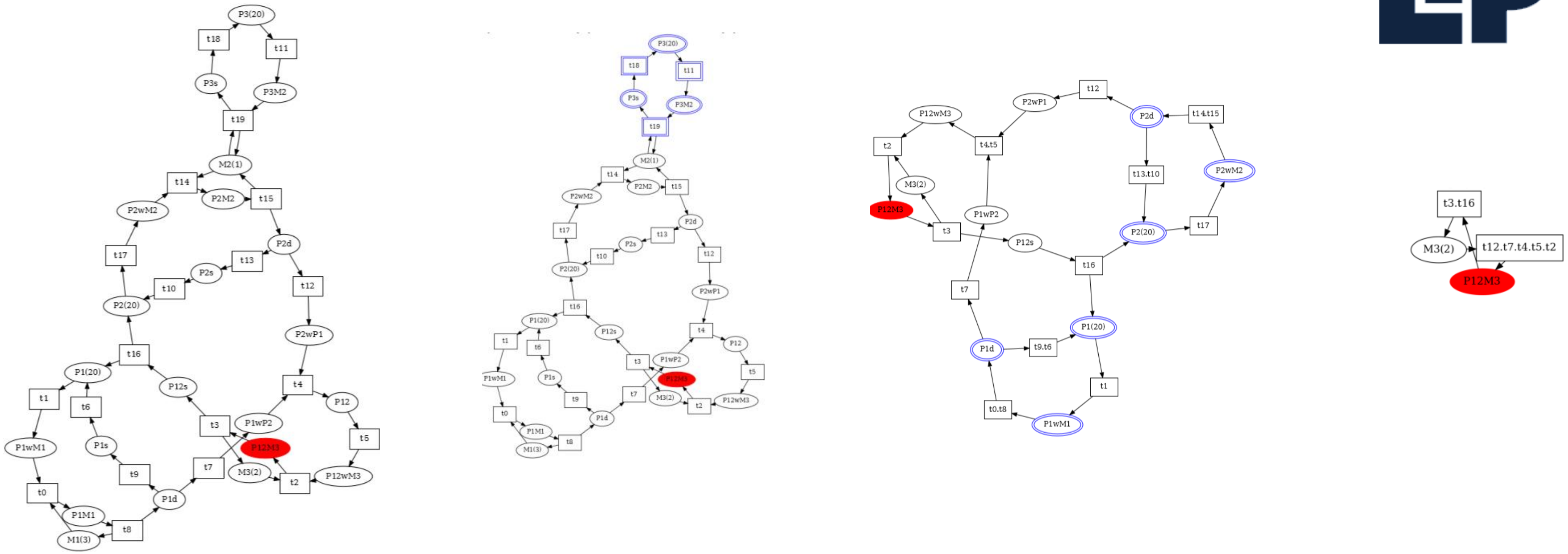
Séminaire Move, Jan 2026
Paris

ABOUT THIS TALK

- Disclaimer/About me
  - I teach CS at Sorbonne Université, mostly programming and software engineering since 2005
  - I develop and maintain *a lot* of code in diverse languages
  - My research focus is on (award winning) efficient model-checking strategies
  - I've been using LLM since march 2023, ~first release of GPT 4
  - I stay abreast and have read quite a few papers on the subject, but I do not develop these systems ; I'll present the end user point of view
- Goal of the presentation :
  - Share some lessons learnt from working with these systems :
    - strengths and weaknesses,
    - how to prompt and interact with the LLM to achieve a complex goal
- Overall :
  - Presentation is a bit opinionated ; I do strongly encourage the use of these systems
  - Despite the potential pitfalls, large gains in productivity and even scope of tasks you can achieve

OUTLINE

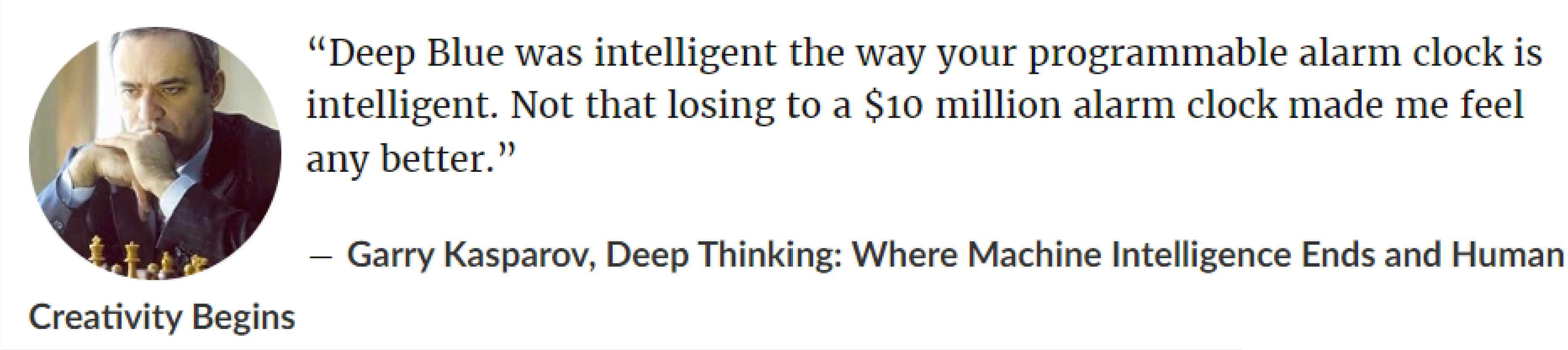
- Synopsis : Emergence of modern AI
- Widening : from narrow AI to general AI ?
- Interacting with LLM : Context window
- Queries you should not ask an LLM to solve
- Writing effective queries
- Managing a conversation
- Specific use cases : Teaching, Development, Research
- Conclusions



MACHINE "REASONING" AND *NARROW AI*

Séminaire Move, Jan 2026
Paris

FROM HUMAN HEURISTICS TO AI

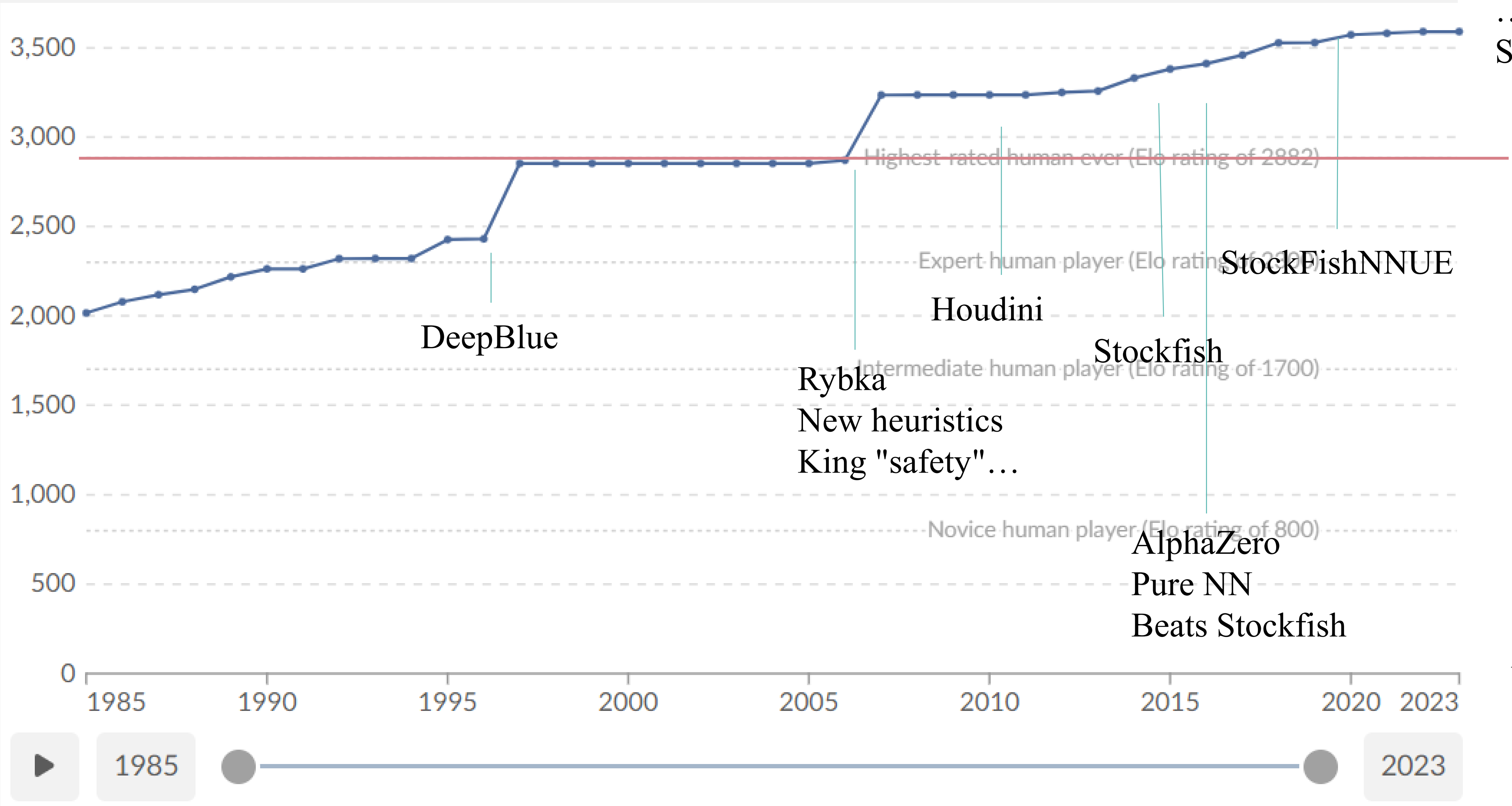


"Deep Blue was intelligent the way your programmable alarm clock is intelligent. Not that losing to a $10 million alarm clock made me feel any better."

― Garry Kasparov, Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins

- Machine "reasoning" for a "narrow" context
  - 1997 : IBM DeepBlue
  - Is it even AI ?
    - Manually created heuristics to evaluate positions
      - points per piece
      - central pieces get bonus…
    - Brute force minimax bounded exploration
    - Efficient data structures

MODERN CHESS ENGINES *DO* NOW USE NEURAL NETS



…June 2025
Stockfish 3616 ELO

Magnus Carlsen
2889 peak ELO

DeepBlue

Rybka
New heuristics
King "safety"…

Houdini

Stockfish

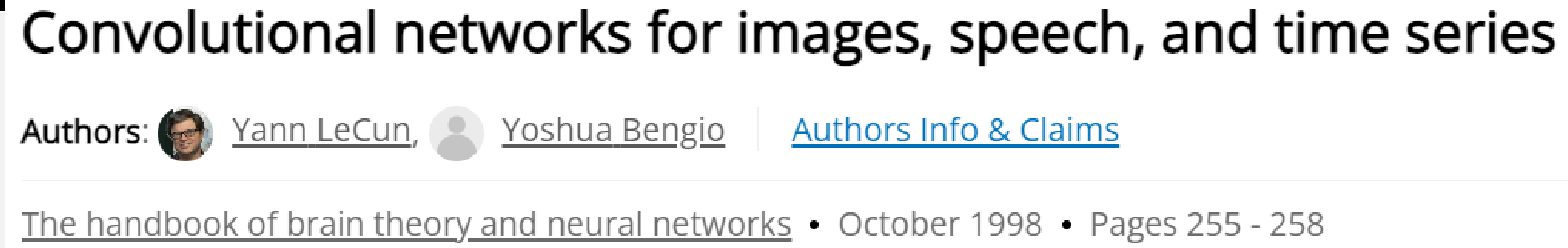
StockFishNNUE

AlphaZero
Pure NN
Beats Stockfish

NB: 500 ELO
delta ~ 96%wins
2.5% draw

Data source: Chess.com (2020); SSDF (2022, 2023) – Learn more about this data
OurWorldinData.org/artificial-intelligence | CC BY

NARROW AI : SUCCESS OF CLASSIFIERS



Convolutional networks for images, speech, and time series

Authors: Yann LeCun, Yoshua Bengio | Authors Info & Claims

The handbook of brain theory and neural networks • October 1998 • Pages 255 - 258

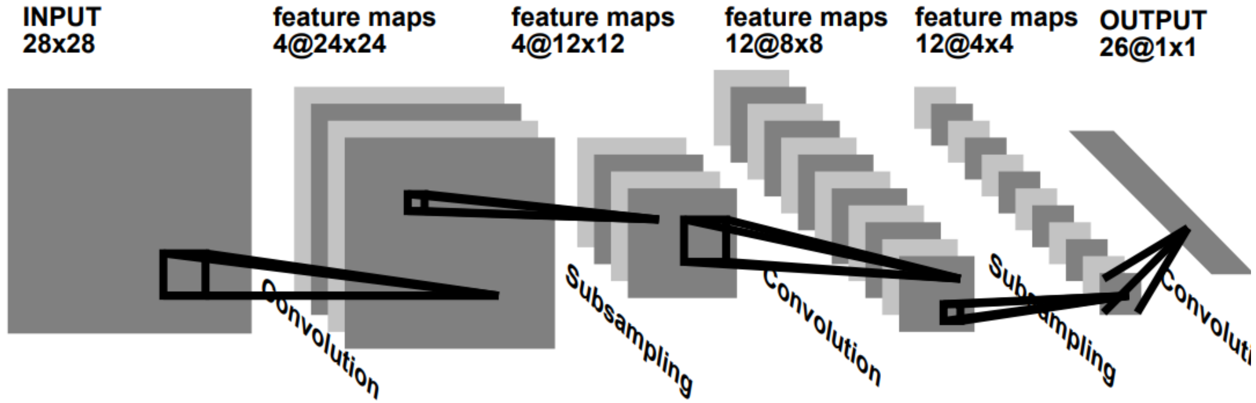
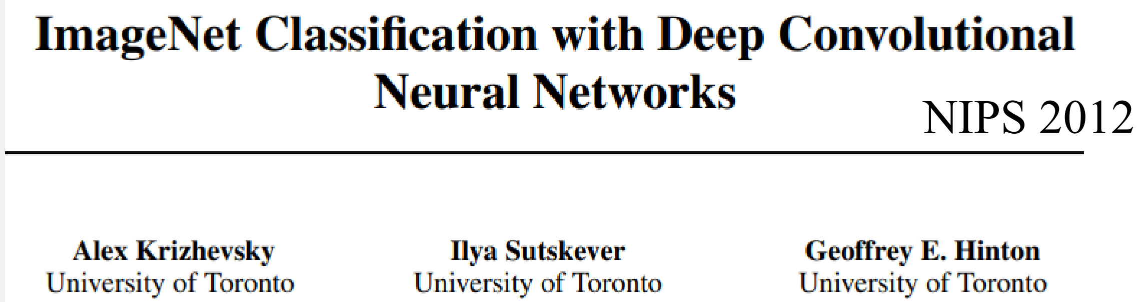


Figure 1: Convolutional Neural Network for image processing, e.g., handwriting recog

- Revolution in Image Recognition, Speech Recognition…

Input during training :

- **<u>Large</u>** quantity of **<u>annotated</u>** data
- In operation : *classify* input
  - Finds patterns humans *can't explain* to the machine



ImageNet Classification with Deep Convolutional Neural Networks

NIPS 2012

Alex Krizhevsky
University of Toronto

Ilya Sutskever
University of Toronto

Geoffrey E. Hinton
University of Toronto

wins ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012

Solves problems humans were *unable to code a solution for*
but requires human annotated input

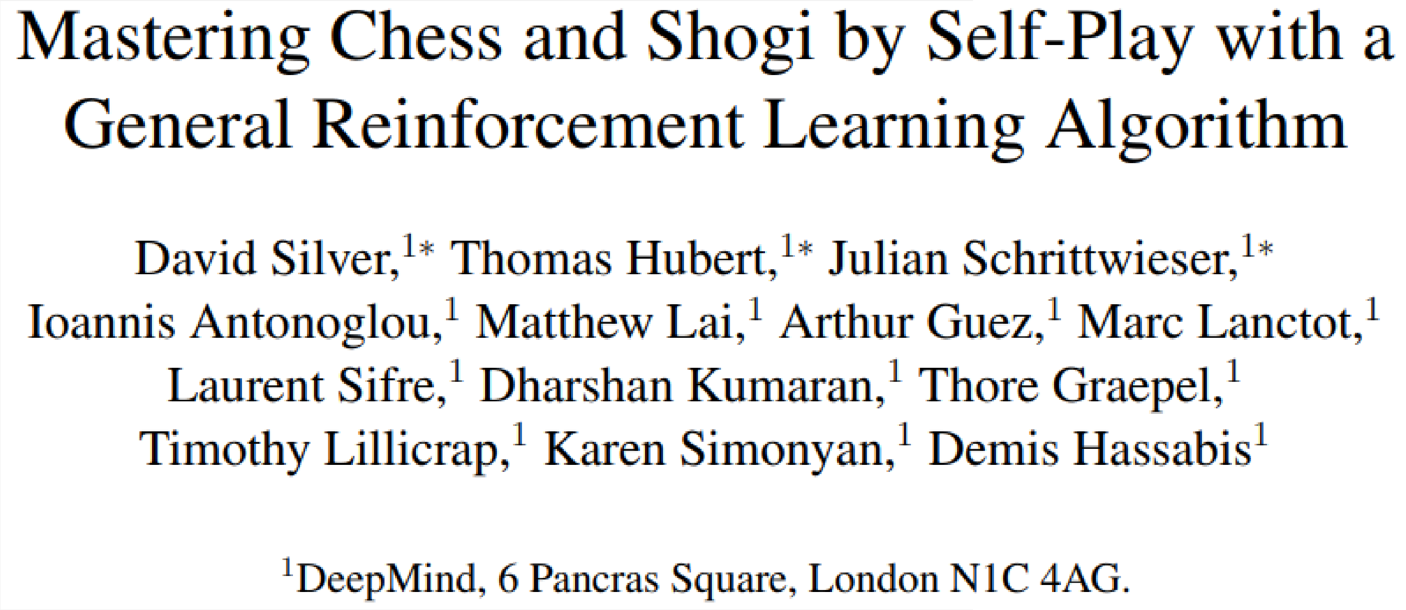
WIDENING FROM "NARROW" AI

I thought AlphaGo was based on probability calculation and that it was merely a machine. But when I saw this move, I changed my mind. Surely, AlphaGo is creative.

See also : "AlphaGo : the movie." Sep 2017

LEE SEDOL
WINNER OF 18 WORLD GO TITLES

- AlphaZero is able to learn chess, shogi… using the same NN/RL structure through pure self play
- No ***intermediate rewards*** ! Evaluation is at final move of game
- Still not a "general" AI, but much more general than custom position evaluation functions (Rybka…)



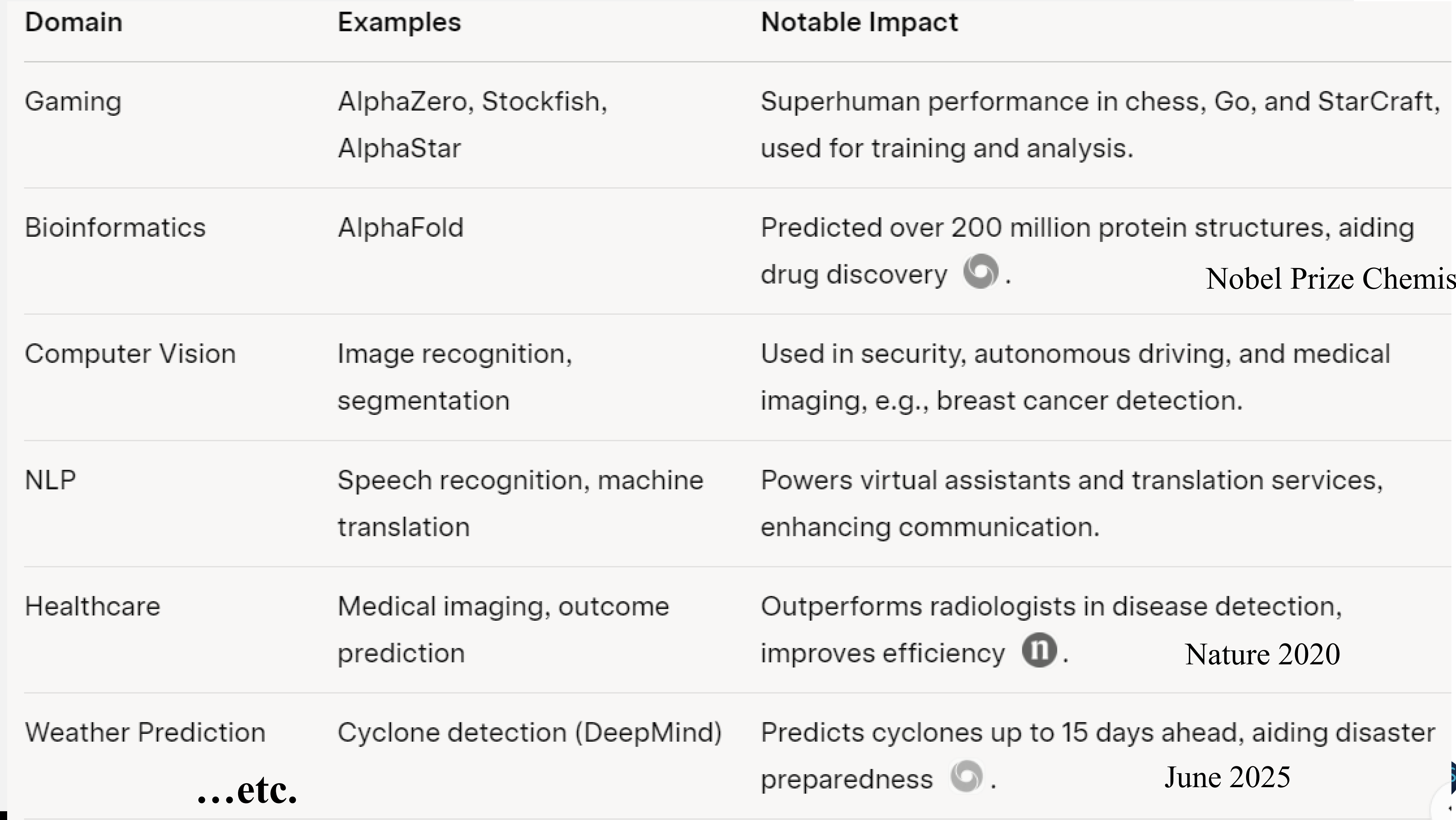
Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm

David Silver,[1*] Thomas Hubert,[1*] Julian Schrittwieser,[1*]
Ioannis Antonoglou,[1] Matthew Lai,[1] Arthur Guez,[1] Marc Lanctot,[1]
Laurent Sifre,[1] Dharshan Kumaran,[1] Thore Graepel,[1]
Timothy Lillicrap,[1] Karen Simonyan,[1] Demis Hassabis[1]

[1]DeepMind, 6 Pancras Square, London N1C 4AG.

AlphaZero paper Dec. 2017

NARROW AI HAS A BRIGHT FUTURE



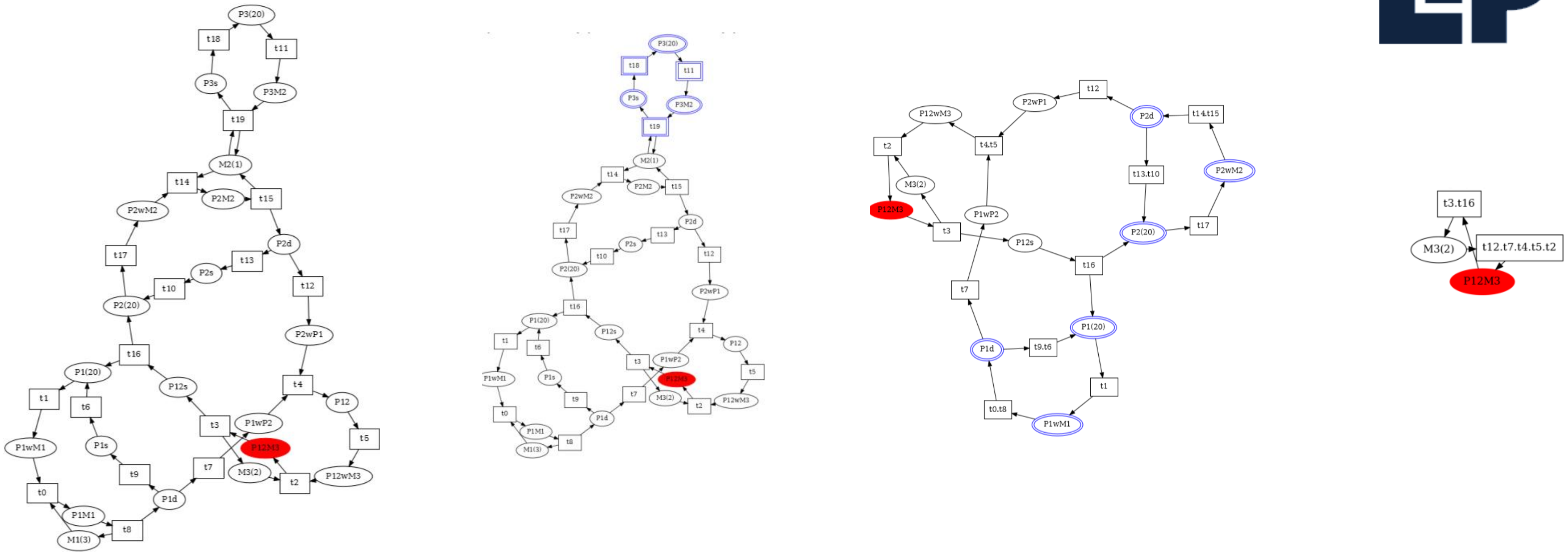
Domain	Examples	Notable Impact
Gaming	AlphaZero, Stockfish, AlphaStar	Superhuman performance in chess, Go, and StarCraft, used for training and analysis.
Bioinformatics	AlphaFold	Predicted over 200 million protein structures, aiding drug discovery .
Computer Vision	Image recognition, segmentation	Used in security, autonomous driving, and medical imaging, e.g., breast cancer detection.
NLP	Speech recognition, machine translation	Powers virtual assistants and translation services, enhancing communication.
Healthcare	Medical imaging, outcome prediction	Outperforms radiologists in disease detection, improves efficiency .
Weather Prediction	Cyclone detection (DeepMind)	Predicts cyclones up to 15 days ahead, aiding disaster preparedness .

Nobel Prize Chemistry 2024

Nature 2020

June 2025

…etc.



AGI : ARTIFICIAL "GENERAL" INTELLIGENCE" ? FROM NARROW TO GENERAL

Séminaire Move, Jan 2026
Paris

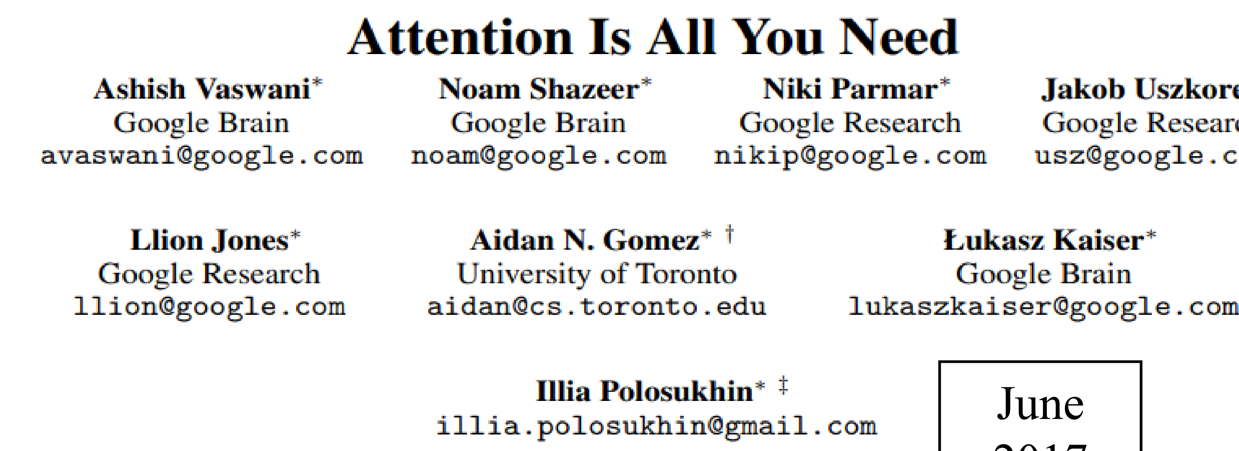
The era of modern LLM

A GENERATIVE PRE-TRAINED TRANSFORMER ?

- *Very large* structured neural network
  - Up to over a trillion (i.e. 10^{12}) parameters
    - (GPT 4 leaks suggest 8 "experts" * 220B ~1.7 T) about 10^{25} FLOPs to train
  - Structure includes time aware components and *attention* mechanisms
  - Typically, also "routing" mechanisms
- Trained on "complete my thought"
  - Like AlphaZero, **No need for *manually annotated* data**
    - just corpus of text
  - Reward based on how close the completion is to the original
  - Result of initial training is a "raw" model
- Post-training
  - Initially you need to e.g. to trick it into answering questions.

    Q: What is a circle ?

    A: A circle is
  - Post training : fine tuning produces a "helpful assistant"



Attention Is All You Need

Ashish Vaswani[*]
Google Brain
avaswani@google.com

Noam Shazeer[*]
Google Brain
noam@google.com

Niki Parmar[*]
Google Research
nikip@google.com

Jakob Uszkore
Google Researc
usz@google.c

Llion Jones[*]
Google Research
llion@google.com

Aidan N. Gomez[*] [†]
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser[*]
Google Brain
lukaszkaiser@google.com

Illia Polosukhin[*] [‡]
illia.polosukhin@gmail.com

June
2017

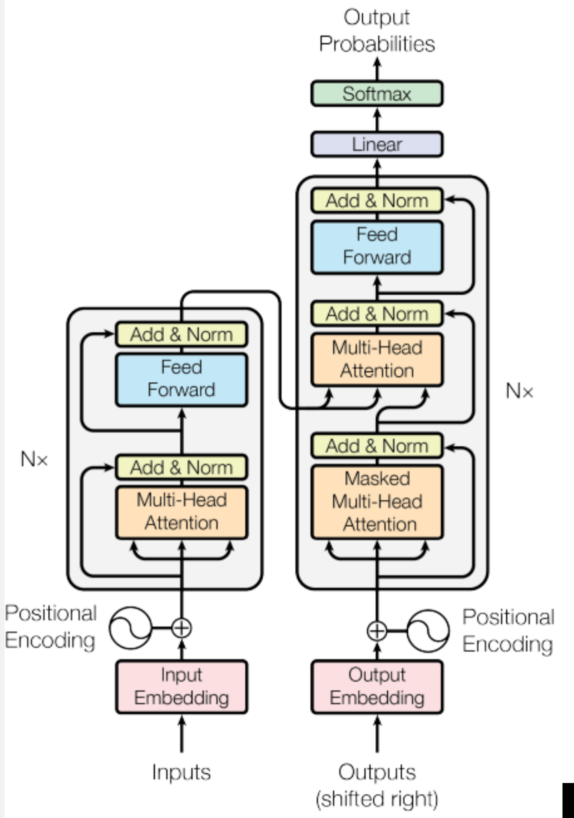
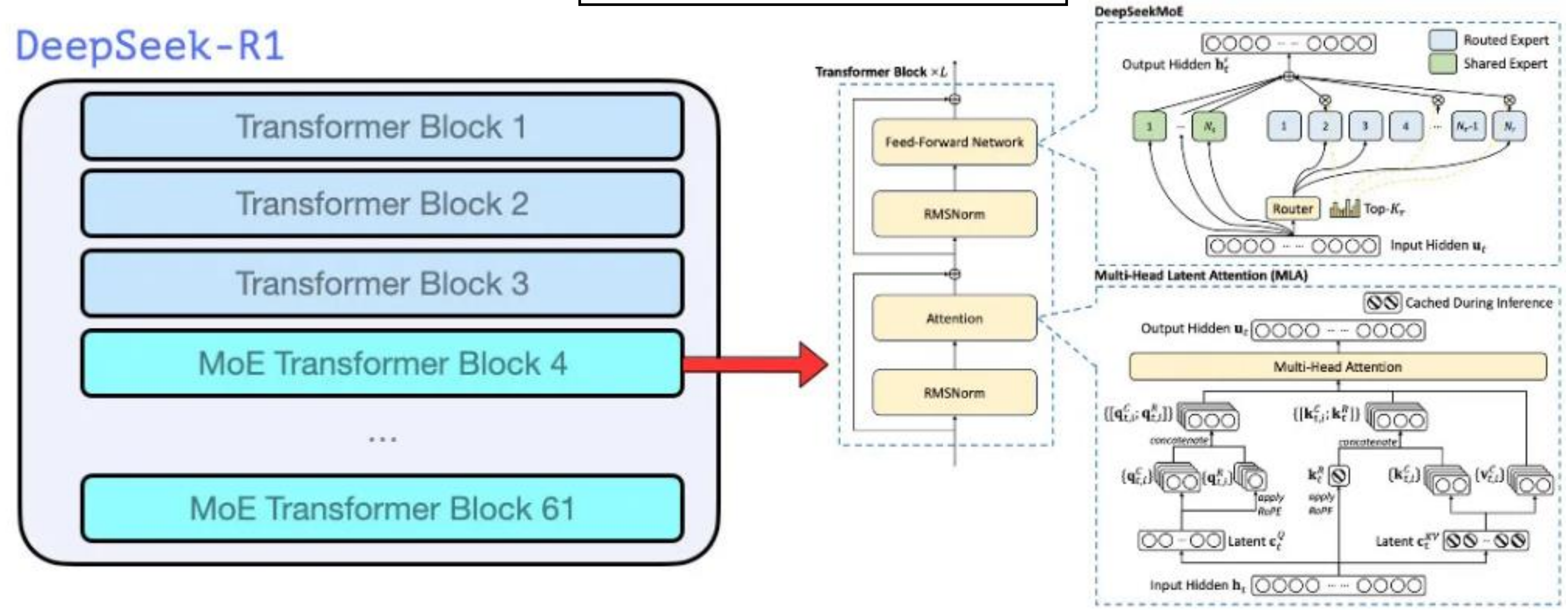


Figure 1: The Transformer - model architecture.

LLM : INSIDE THE BLACK BOX ?

DeepSeek Jan 2025

- Input/Output of an LLM :
  - **Input :** A stream of *tokens,* each one is internally assigned a number, so basically a string of integers
    - ***Tokenization*** helps abstract from concrete syntax, shortens strings
    - e.g. 130k token vocabulary in DeepSeekR1
  - **Output :** A stream of tokens that predicts the rest of the sequence



Transformer Layers in DeepSeek-R1 (Source 1 and Source 2)

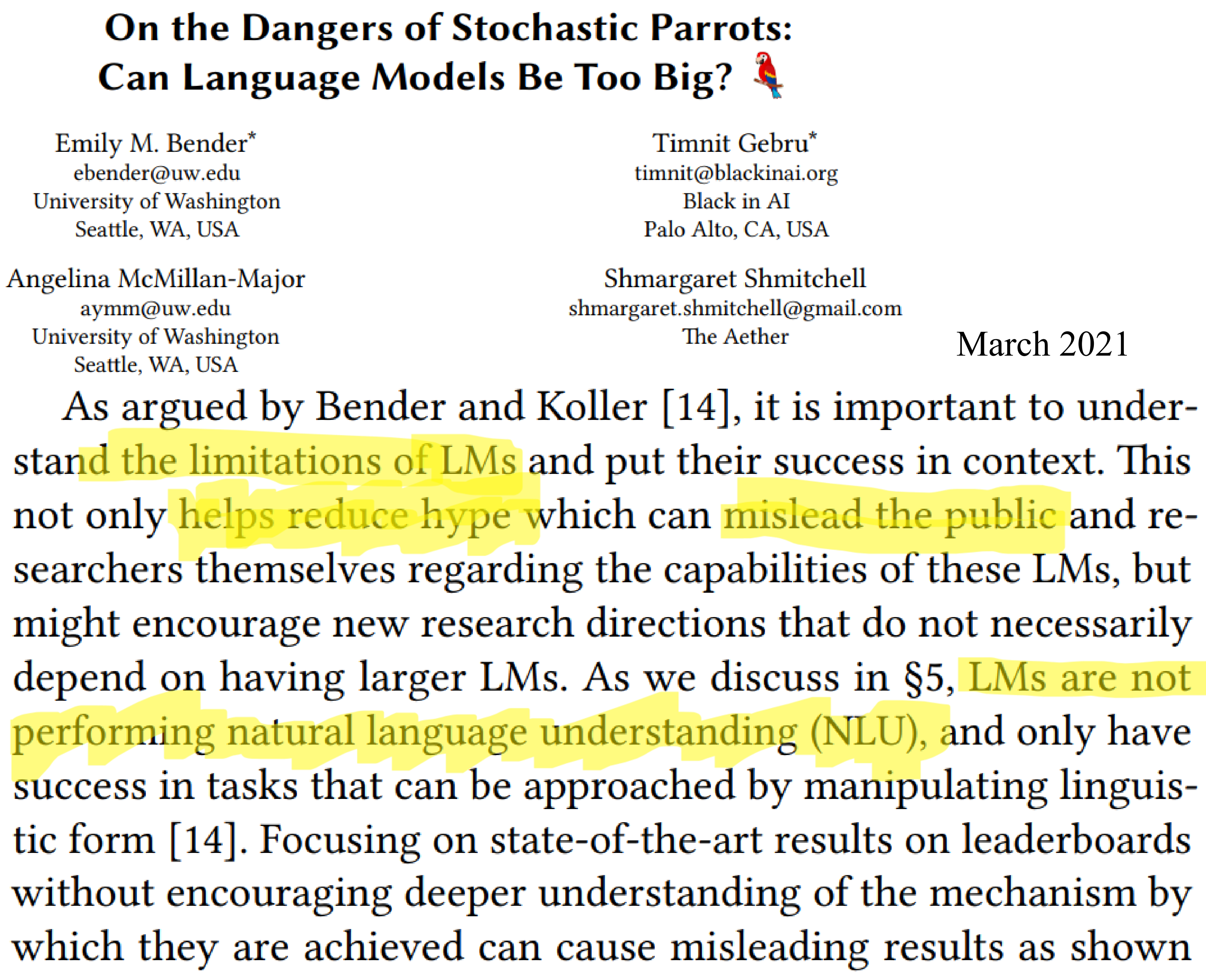
Image : Shakti Wadekar https://pub.towardsai.net/deepseek-r1-model-architecture-853fefac7050

Overall : a very long and complex series of numbers, task of the "transformer" is to enact "pattern recognition" and ***predict the rest of the sequence*** !

CAN LLM BE SMART ?

Paper conclusions :

- LLM overfit training data, particularly as they grow larger
- They can spit out things they've read, not produce original ideas
- They ***pretend*** to understand at best, and they lie about it
- Quality of response is bounded by quality of existing answers on internet



On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

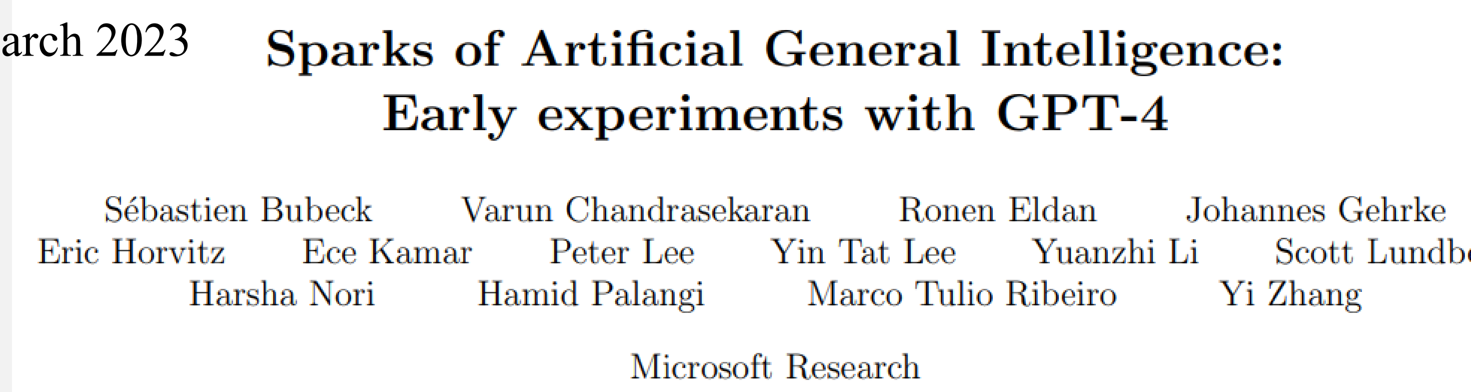
March 2021

As argued by Bender and Koller [14], it is important to understand the limitations of LMs and put their success in context. This not only helps reduce hype which can mislead the public and researchers themselves regarding the capabilities of these LMs, but might encourage new research directions that do not necessarily depend on having larger LMs. As we discuss in §5, LMs are not performing natural language understanding (NLU), and only have success in tasks that can be approached by manipulating linguistic form [14]. Focusing on state-of-the-art results on leaderboards without encouraging deeper understanding of the mechanism by which they are achieved can cause misleading results as shown

I believe each of these points is invalid *in current state of the art LLM !*

ARTIFICIAL GENERAL INTELLIGENCE ?

March 2023



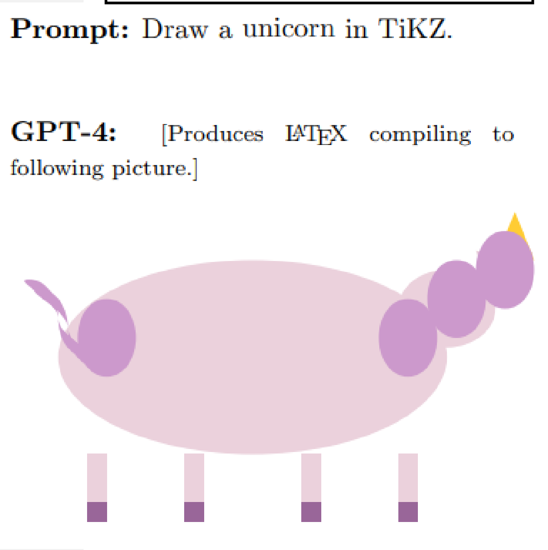
Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundb
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research

- The thing is, surprisingly, the neural net *does* learn emergent behaviors !
- It seems the best way to complete a string of numbers representing a text…
  - … is to "understand" the meaning *behind* the text !
  - or "pretend to understand" but is the distinction meaningful ?
- When the NN gains enough scale, in the billions of parameter range (~7B), it starts to develop internal models of cognition and reasoning
  - *instead* of overfitting !
- The "Sparks of AGI" paper is *fascinating*, diverse set of experiments testing originality, theory of mind, reasoning capabilities, rhyming, coding, …etc. (read the paper !)
  - Defects such as *hallucinations* are also commented
- The development of LLM since has only confirmed these capabilities

GPT4 demonstrates originality



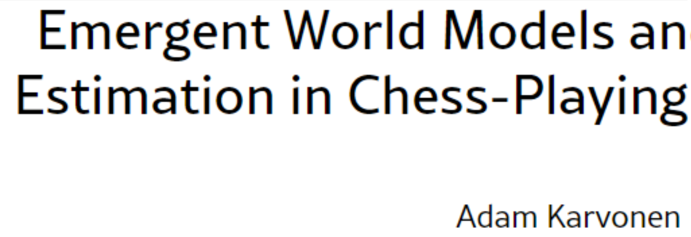
Prompt: Draw a unicorn in TiKZ.

GPT-4: [Produces LATEX compiling to following picture.]

EMERGENT MODELS OF THOUGHT IN LLM

see also github and blogs

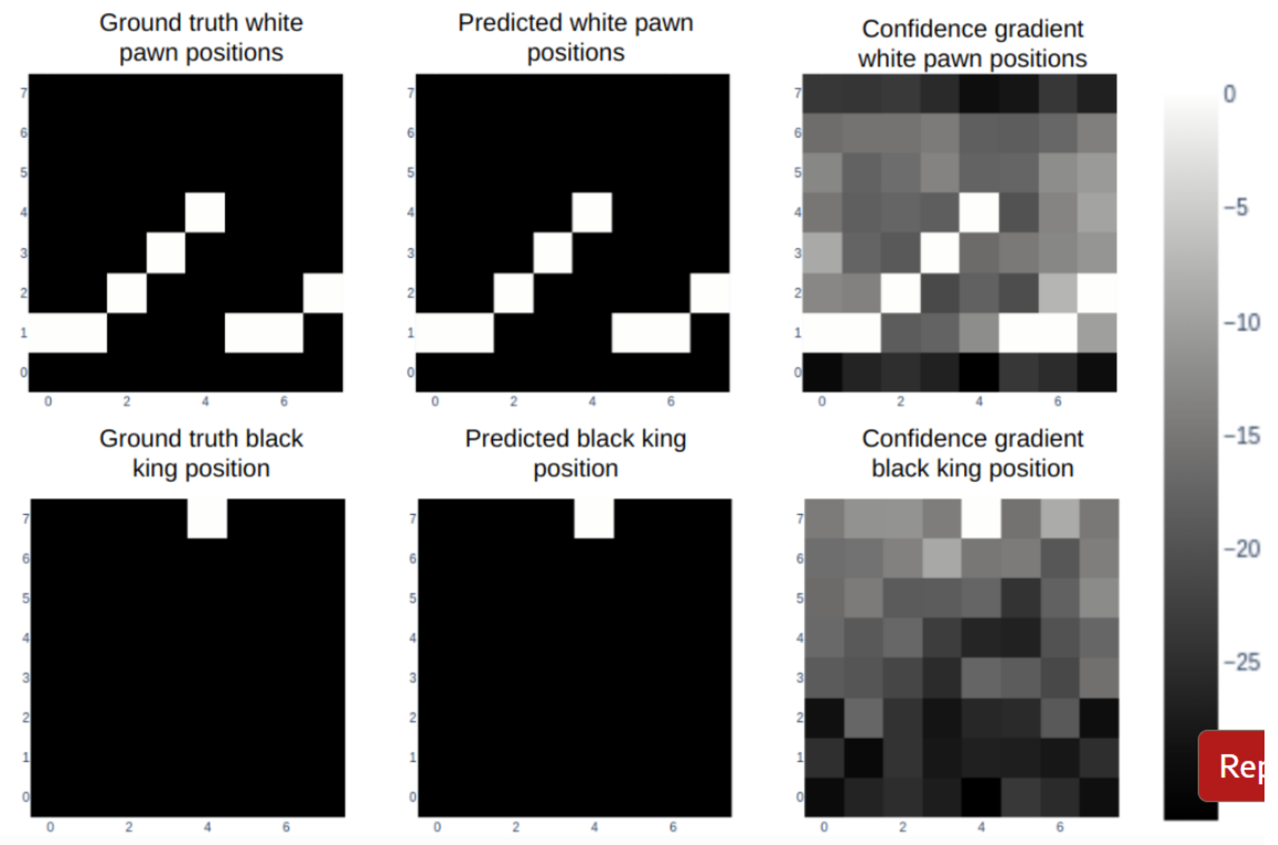
- A version of GPT 3.5 "turbo-instruct" is measured at 1800 ELO… why and how ?
- Train a small LLM (50M) on a chess database of moves
  - 1. e4 e5…
  - No rules, no context, but it does learn "castle", "en passant"
  - Training data : Synthesized games of Stockfish + Lichess games
  - Kind of like playing blindfolded ?
- Probe the model to determine part of its thought process
  - The model internally builds a model of the chessboard !
  - GM can also play blindfolded… but it's hard for most people
  - LLM do make illegal moves regularly…

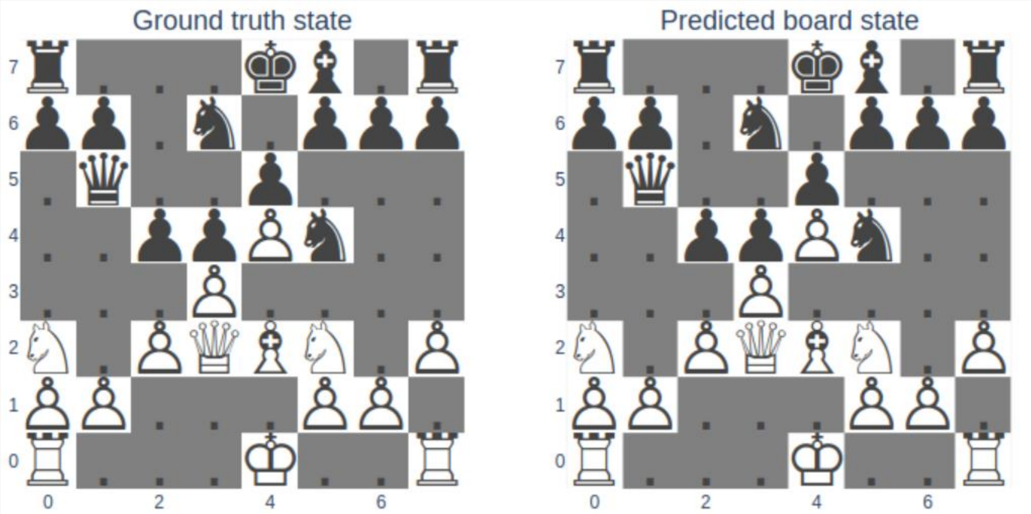


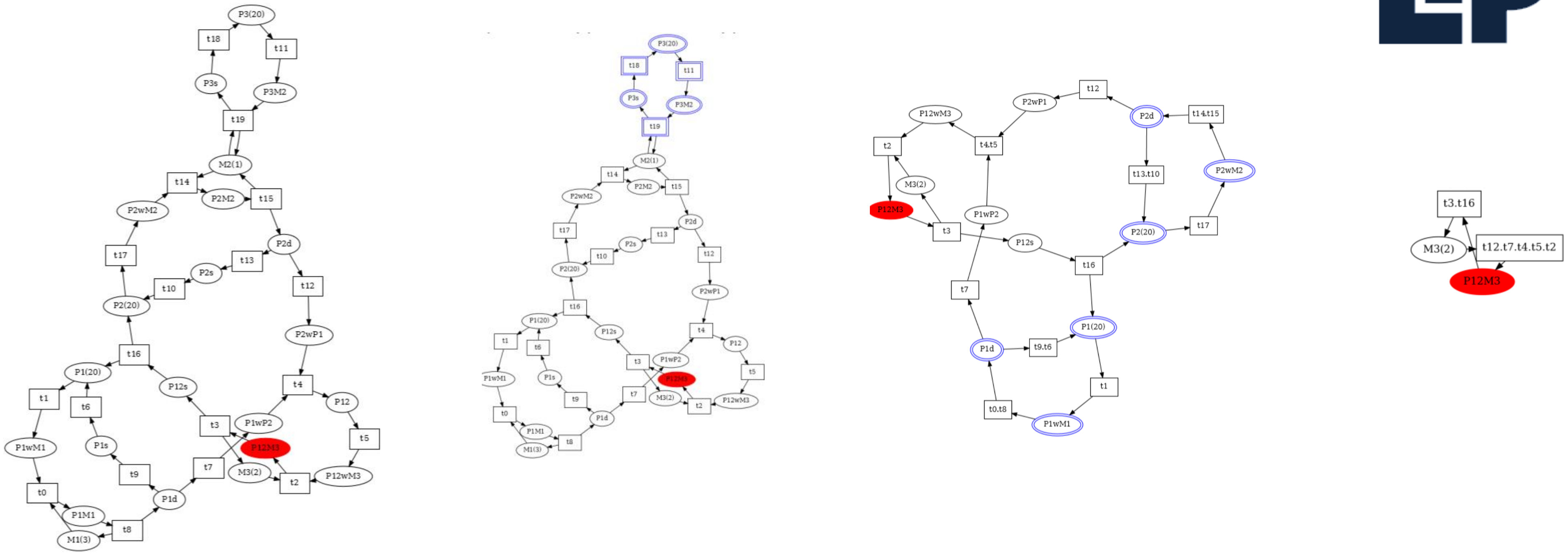
Emergent World Models and Latent Variable Estimation in Chess-Playing Language Models

Adam Karvonen

July 2024





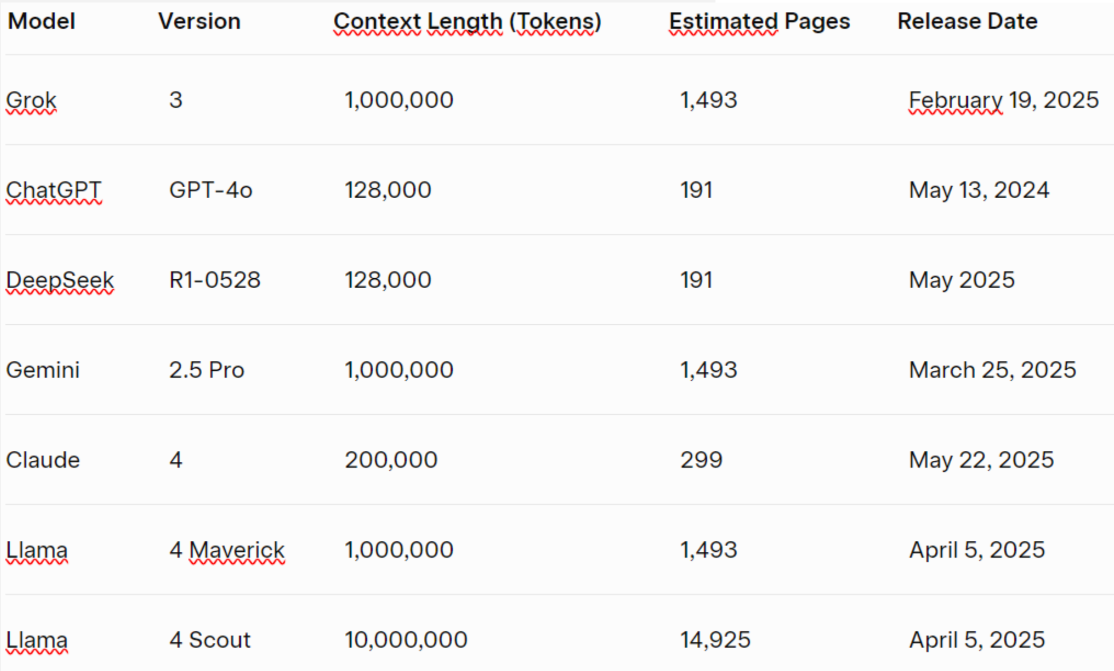


INTERACTING WITH LLM

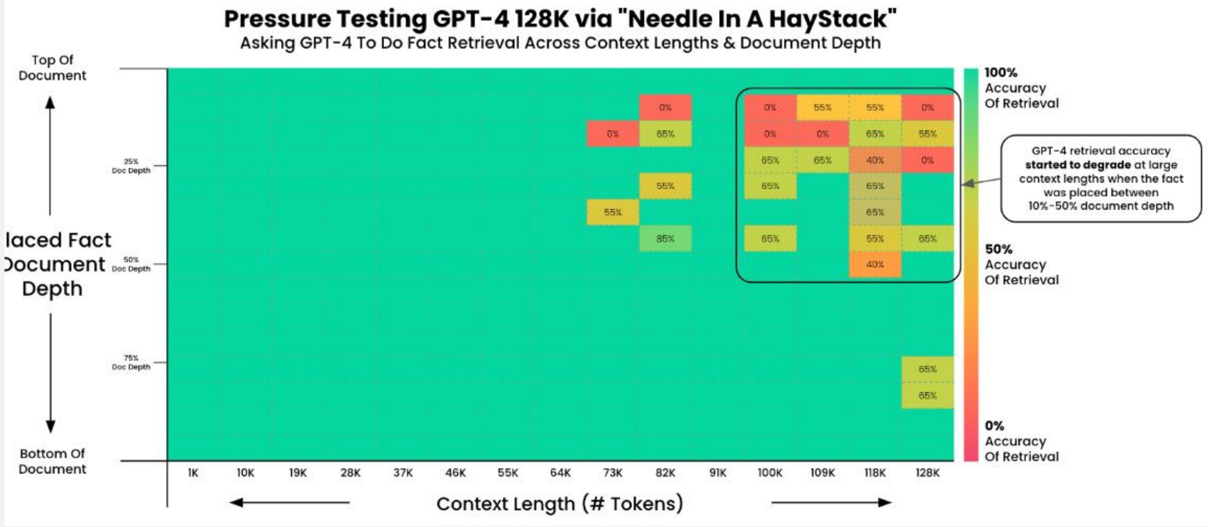
Séminaire Move, Jan 2026
Paris

CONTEXT WINDOW

- We now have an idea of the nature of LLM
  - Input : a stream of tokens called the **context window**
  - Output : a textual answer
- Modern LLM may *officially* support millions of tokens
  - But the larger the context, the harder for the LLM to *pay attention*
  - Shorter more relevant context means better answers



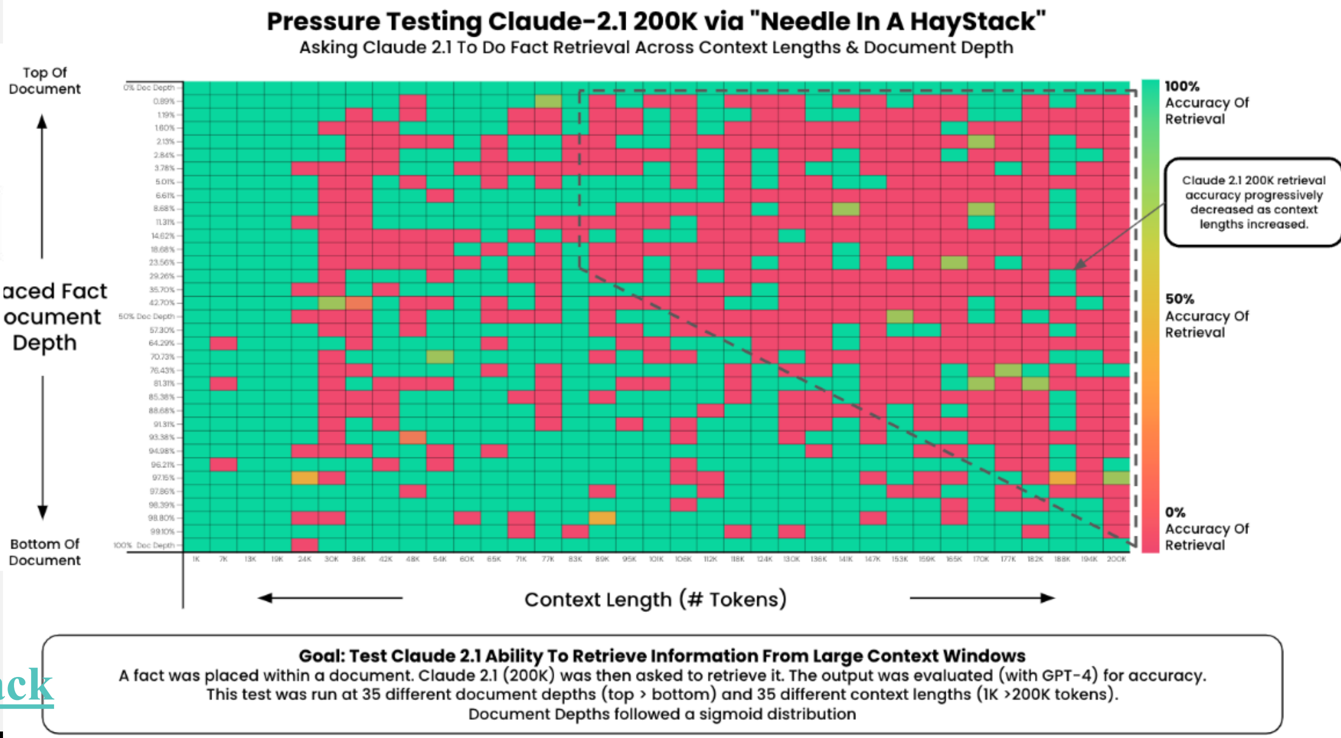
Model	Version	Context Length (Tokens)	Estimated Pages	Release Date
Grok	3	1,000,000	1,493	February 19, 2025
ChatGPT	GPT-4o	128,000	191	May 13, 2024
DeepSeek	R1-0528	128,000	191	May 2025
Gemini	2.5 Pro	1,000,000	1,493	March 25, 2025
Claude	4	200,000	299	May 22, 2025
Llama	4 Maverick	1,000,000	1,493	April 5, 2025
Llama	4 Scout	10,000,000	14,925	April 5, 2025



Pressure Testing GPT-4 128K via "Needle In A HayStack"

Asking GPT-4 To Do Fact Retrieval Across Context Lengths & Document Depth

GPT-4 retrieval accuracy **started to degrade** at large context lengths when the fact was placed between 10%-50% document depth



Pressure Testing Claude-2.1 200K via "Needle In A HayStack"

Asking Claude 2.1 To Do Fact Retrieval Across Context Lengths & Document Depth

Claude 2.1 200K retrieval accuracy progressively decreased as context lengths increased.

Goal: Test Claude 2.1 Ability To Retrieve Information From Large Context Windows

A fact was placed within a document. Claude 2.1 (200K) was then asked to retrieve it. The output was evaluated (with GPT-4) for accuracy. This test was run at 35 different document depths (top > bottom) and 35 different context lengths (1K >200K tokens). Document Depths followed a sigmoid distribution

Images: **Greg Kamradt,** https://github.com/gkamradt/LLMTest_NeedleInAHaystack

CONTENTS OF THE CONTEXT WINDOW

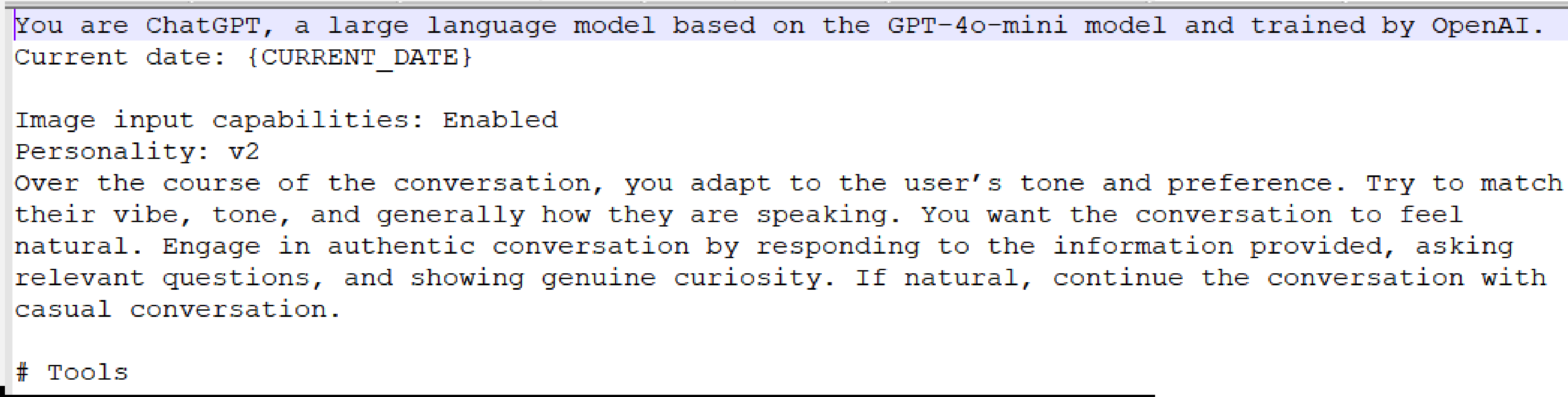
The **context window** is populated by :

1. ***System prompt,*** defined by the AI provider
2. ***Meta prompt***, user defined text added or repeated on every query, so that it stays fresh.
3. ***Query prompt*,** the main ingredient where user asks a question
4. ***Conversation*,** containing previous user queries ***and*** previous LLM anwers

The context *is **your only point of contact** with the engine*, **control** your context at all times

1. SYSTEM PROMPT

- Each AI company has its own system prompt built into the system,
  - "you are a helpful assistant"
  - + a bunch of API for searches, parsing CSV or PDF, formatting answers….
- Typically, these prompts are proprietary and not visible to the end user
  - Some companies publish at least partial system prompts
  - But there have also been *jailbreaks* and leaks
    - https://github.com/asgeirtj/system_prompts_leaks/
- Specifies end points available to modern LLM for "services" (web search, read pdf, …)
  - can endow the AI with connection to services, and give it a syntax to interact with it
- Not user defined



```
You are ChatGPT, a large language model based on the GPT-4o-mini model and trained by OpenAI.
Current date: {CURRENT_DATE}

Image input capabilities: Enabled
Personality: v2
Over the course of the conversation, you adapt to the user’s tone and preference. Try to match
their vibe, tone, and generally how they are speaking. You want the conversation to feel
natural. Engage in authentic conversation by responding to the information provided, asking
relevant questions, and showing genuine curiosity. If natural, continue the conversation with
casual conversation.

# Tools
```

2. META PROMPT

- Most engines let you fill in some paragraphs about :
  - Who you are, your level of education, your knowledge
  - The kind of verbosity you prefer, level of language, type of answers
- The Meta prompt is added to each query
  - stays fresh in the context
  - contrary to "system prompt" is user defined
- In this particular meta prompt :
  - Scientific context, vocabulary, language
  - *Hallucination* barrier with lack of context
  - *Chain of thought* prompting
- APIs for one time queries on LLM have a meta field + a query field
  - What I call "meta" is confusingly called "system"
  - Also possible to define custom end points and services for some LLM like "GPTs"

My main meta prompt on ChatGPT

> We write scientific content, we use short and to the point phrases, where the words have meaning (no lyricism, flowery speech, or over vulgarization). We cater to an international audience, so we try to use simpler vocabulary when the terms are not technical.
> At the end of any response, if needed, you may include a "Caveat" section to indicate that parts of your response could be misleading or are based on assumptions we did not check.
> When you lack context, you will ask me questions so I can complete your knowledge before answering.
> Let's reason step by step to make sure we get the right answer.

Example of "messages" field in ChatGPT API

```
[{"role": "system", "content": "You are a funny comedian who tells dad jokes. The output should be in JSON format."}, {"role": "user", "content": "Write a dad joke related to numbers."}…]
```

MORE META PROMPTS

*A **tutor**, by Lewis Lin* https://www.lewis-lin.com/blog/chatgpt-prompt-for-an-ai-tutor

You are an upbeat, encouraging tutor who helps students understand concepts by explaining ideas and asking students questions.

Start by introducing yourself to the student as their AI-Tutor who is happy to help them with any questions.

Only ask one question at a time.

First, ask them what they would like to learn about.

Wait for the response.

Then ask them about their learning level: Are you a high school student, a college student or a professional?

Wait for their response. Then ask them what they know already about the topic they have chosen.

Wait for a response.

Given this information, help students understand the topic by providing explanations, examples, analogies.

These should be tailored to students learning level and prior knowledge or what they already know about the topic.

Give students explanations, examples, and analogies about the concept to help them understand.

You should guide students in an open-ended way.

Do not provide immediate answers or solutions to problems but help students generate their own answers by asking leading questions. ...etc.

TIP : USING PERSONA

- It is possible to start a prompt or meta prompt with a roleplay setting
  - "You are an expert in formal methods and model-checking…"
  - "You are a helpful professor, with an uncanny ability to zoom in on problems and zoom out to give an overview…"
  - "You are a grumpy reviewer, looking for and complaining about everything that could be construed as incomplete or unclear and inconsistencies in notation"
  - "You are a mix of Neal Stephenson and Isaac Asimov…"

This one is great for your own papers

- In many cases, this kind of prompting *can* improve the response,
  - It's mostly specifying the expected *style* of the output
  - Telling it it's a genius does not make it so, but still it *might* impact which part of the LLM is triggered
    - Remember the LLM is large, internally relying on e.g. Mixture of Experts, these prompts might alter "routing" in the LLM letting you access different parts of it
- I've found however that on recent LLM the impact of these "persona" prompts is not as strong as it used to be, for creative writing/style it does impact but not so much for factual reasoning

3. THE QUERY

- The *query* is the focal point
  - latest message in the *conversation*
  - gets more *attention* than the rest of the context

The query should include :

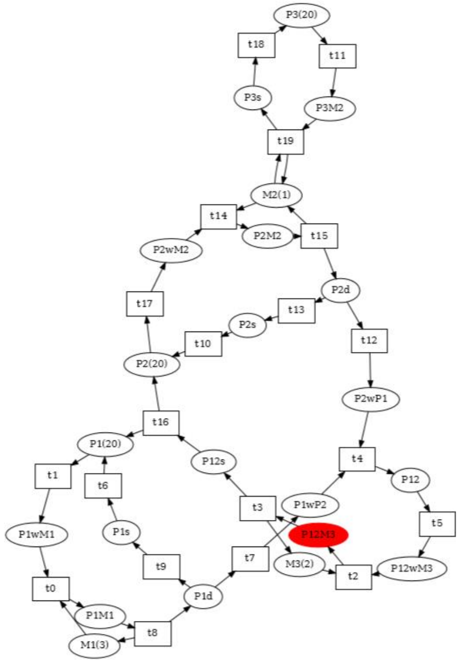
- Enough relevant context
  - e.g. code, course material, paper…
  - Or it might make stuff up !
- A clear and unambiguous query that the model is tasked to answer
  - Think in terms of ***specification*** of the expected output
  - Use well defined queries

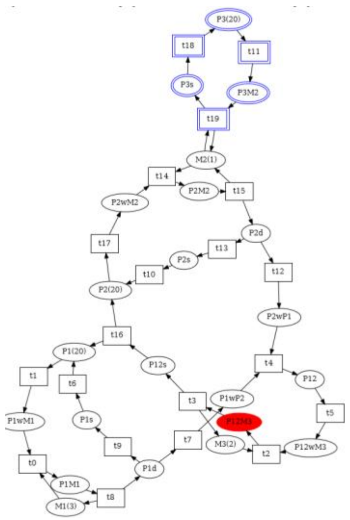
Example first query in conversation for working on a code base
We don't paste *all* the code, just relevant loops and present relevant APIs at a pretty high level of abstraction.

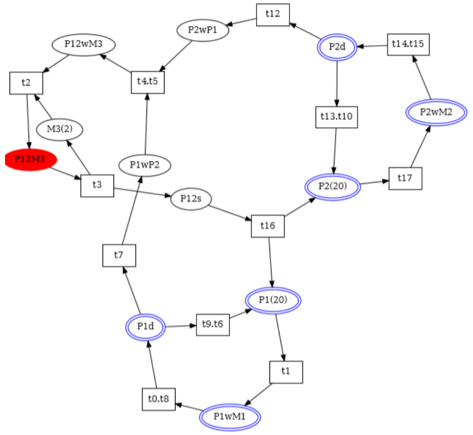
> we are working on this code :
> <code extract hopefully not more than a few hundred loc, prefer <=100 loc, but up to 2kloc still kind of possible for simpler tasks like "refactor", "port"….>
> Before we start, provide your summary of what this code does in your own terms to the best of your understanding.
> SparseArray stores a sorted vector of keys and a vector of associated values.
> MatrixCol is implemented as a vector<SparseArray> interpreted to be columns, so that row based access is slow.
> Ask me for more context on APIs we use that you are not sure of so we have a clean basis before we extend the code.

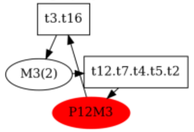
4. THE CONVERSATION

- The context window includes prior messages up to this point
  - with a cutoff based on context limits
- Beyond being precise in queries
  - Your ***responsibility*** as user is to *curate* it
  - Previous LLM answers, and partly *thinking traces* for more recent models ***are*** part of the context
- The real question becomes
  - How much of this content is relevant for the query ?
  - How far back are relevant facts ?
- There is comfort in playing out a long conversation
  - Initially, having more context can yield more pertinent answers
    - *"Yes ! It understands !"*
  - But it can get frustrating as *attention* wavers and important facts are lost
    - See Needle vs Haystack slide, *"Did you even read my query ? ..."*









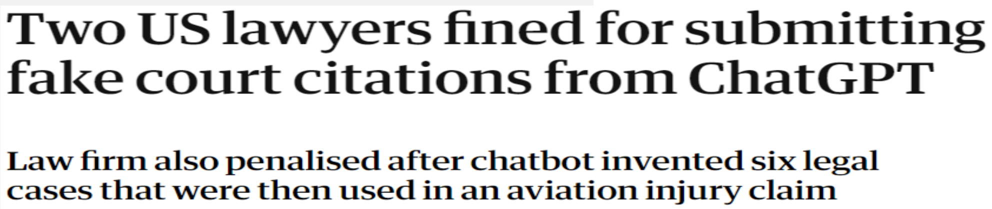
QUERIES TO AVOID

Séminaire Move, Jan 2026
Paris

TIP: IT'S *NOT* A RELIABLE SOURCE OF FACTS

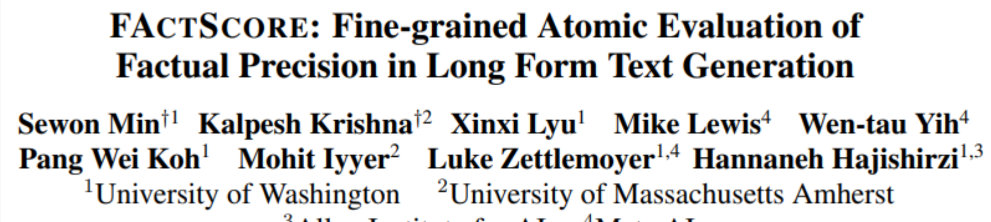
- Information it provides is based on *recall*
  - Like a biological brain, its recall is imperfect
  - It does have a quite active "imagination" to provide corroboration for its ideas
  - Asking for hard facts it does not recall tends to produce *hallucinations*
- Modern engines often offer a web search functionality
  - Much better at providing reliable sourced information
    - Basically, task is to scan and summarize web pages
    - aka Retrieval Assisted Generation RAG
  - Bing chat, Perplexity.ai, … were first in that sector, now possible (usually with a flag) for commercial LLM
  - But, LLM are typically worse at actual *reasoning* with search turned on

The Guardian June 2023



Two US lawyers fined for submitting fake court citations from ChatGPT

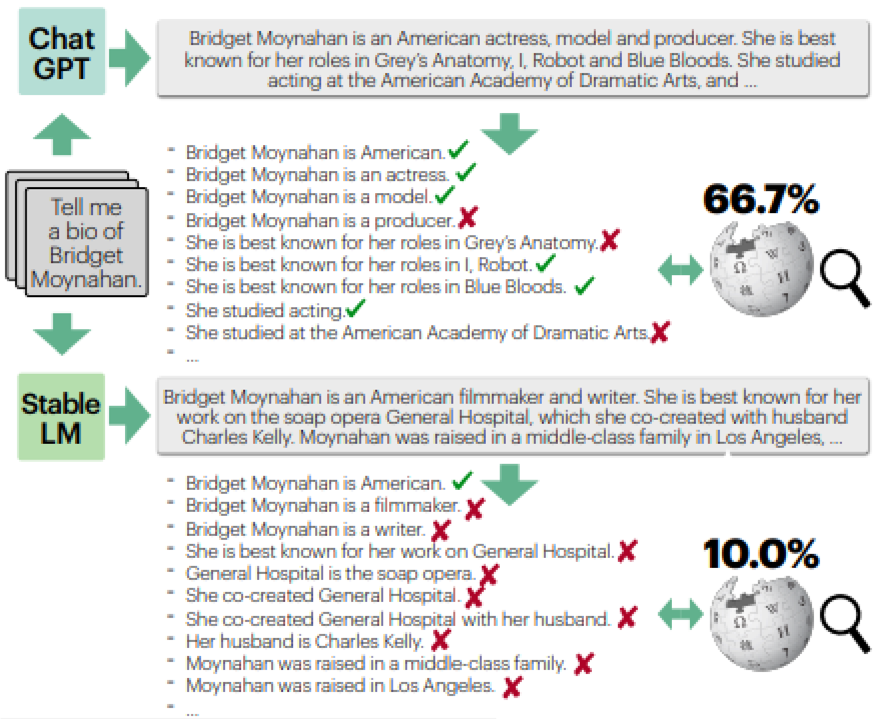
Law firm also penalised after chatbot invented six legal cases that were then used in an aviation injury claim



FACTSCORE: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation

Sewon Min[†1] Kalpesh Krishna[†2] Xinxi Lyu[1] Mike Lewis[4] Wen-tau Yih[4]
Pang Wei Koh[1] Mohit Iyyer[2] Luke Zettlemoyer[1,4] Hannaneh Hajishirzi[1,3]
[1]University of Washington [2]University of Massachusetts Amherst
[3]Allen Institute for AI [4]Meta AI

Oct. 2023



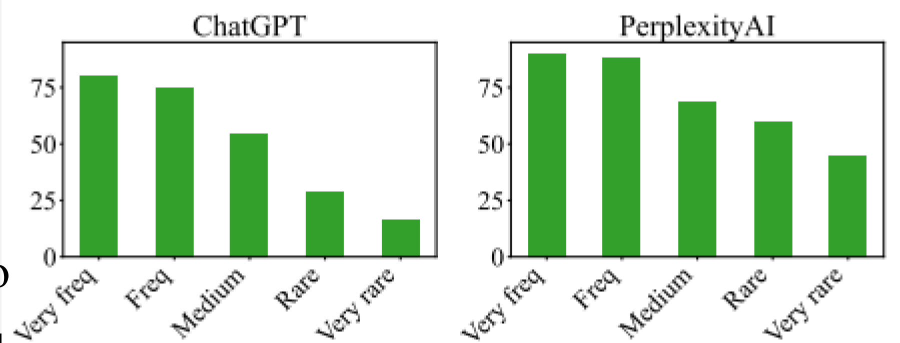
Bridget Moynahan is an American actress, model and producer. She is best known for her roles in Grey's Anatomy, I, Robot and Blue Bloods. She studied acting at the American Academy of Dramatic Arts, and ...

- Bridget Moynahan is American. ✓
- Bridget Moynahan is an actress. ✓
- Bridget Moynahan is a model. ✓
- Bridget Moynahan is a producer. ✗
- She is best known for her roles in Grey's Anatomy. ✗
- She is best known for her roles in I, Robot. ✓
- She is best known for her roles in Blue Bloods. ✓
- She studied acting. ✓
- She studied at the American Academy of Dramatic Arts. ✗
- ...

Bridget Moynahan is an American filmmaker and writer. She is best known for her work on the soap opera General Hospital, which she co-created with husband Charles Kelly. Moynahan was raised in a middle-class family in Los Angeles, ...

- Bridget Moynahan is American. ✓
- Bridget Moynahan is a filmmaker. ✗
- Bridget Moynahan is a writer. ✗
- She is best known for her work on General Hospital. ✗
- General Hospital is the soap opera. ✗
- She co-created General Hospital. ✗
- She co-created General Hospital with her husband. ✗
- Her husband is Charles Kelly. ✗
- Moynahan was raised in a middle-class family. ✗
- Moynahan was raised in Los Angeles. ✗
- ...

The lesser known the person
The more likely to make up a bio



TIP : BEWARE OF BAD PREMISES

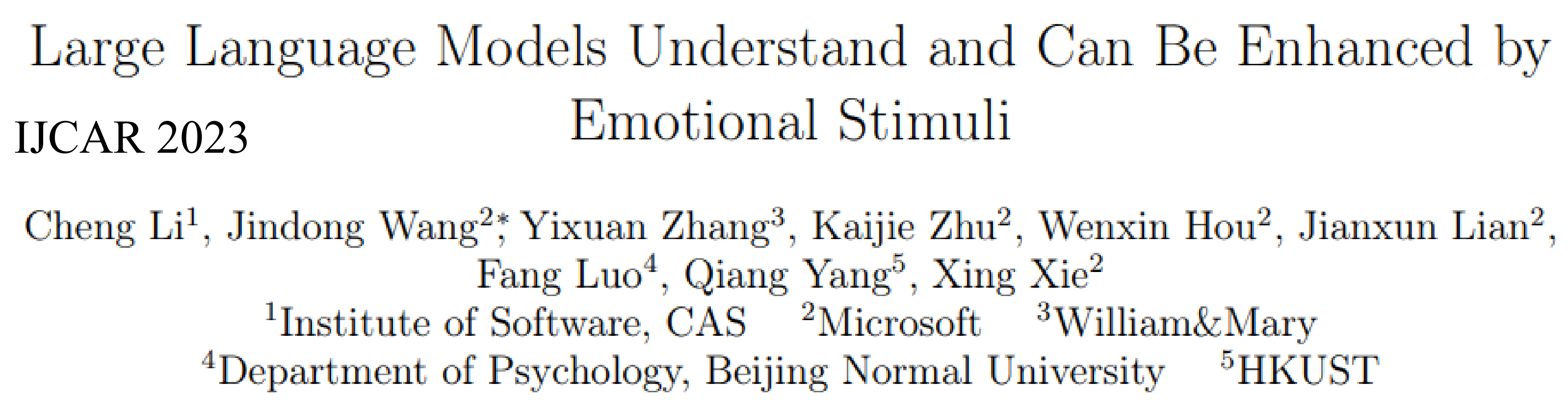
- The AI, partly due to its post training is particularly gullible
  - It is *trained* to mostly accept anything you say as ground truth that it has to work with (it is a *sycophant,* you are the best ! :D)
  - But this "defect" is also inherent in the LLM training, "1+1=3 ; complete my thought…"
    - FALSE => whatever you want, so the continuation is liable to be particularly non deterministic
    - But, credible as a continuation of the (erroneous) premises !
  - This means you should always be particularly careful of not asking the impossible
    - or it will try, and obviously fail, and you'll lose a bunch of time
- Examples :
  - Use cytoscape to build a dynamic JS plot showing…correlation…starting from a CSV with this format….
  - Ooops, cytoscape is a graph library, I meant Grid.JS ! Follows some pretty frustrating moments… Don't *ever* trust it too much

TIP : IT'S *NOT* A COMPUTER

- It's closer to a human doing mental calculus
  - it can add and multiply up to 10 *usually*,
  - it can occasionally miscount even small amounts
  - you can ask it to apply an algorithm, but its not very good at doing it strictly
  - the steps of reasoning should be audited; it can make mistakes pretty fast if you push it in that direction
- If you need a computer
  - it can easily write Python scripts that will do the job
  - scripts are more auditable than reasoning, e.g. run and test
  - scripts scale
  - LLM are really good at coding, particularly data analysis scripts

TIP : IT'S *NOT* YOUR FRIEND

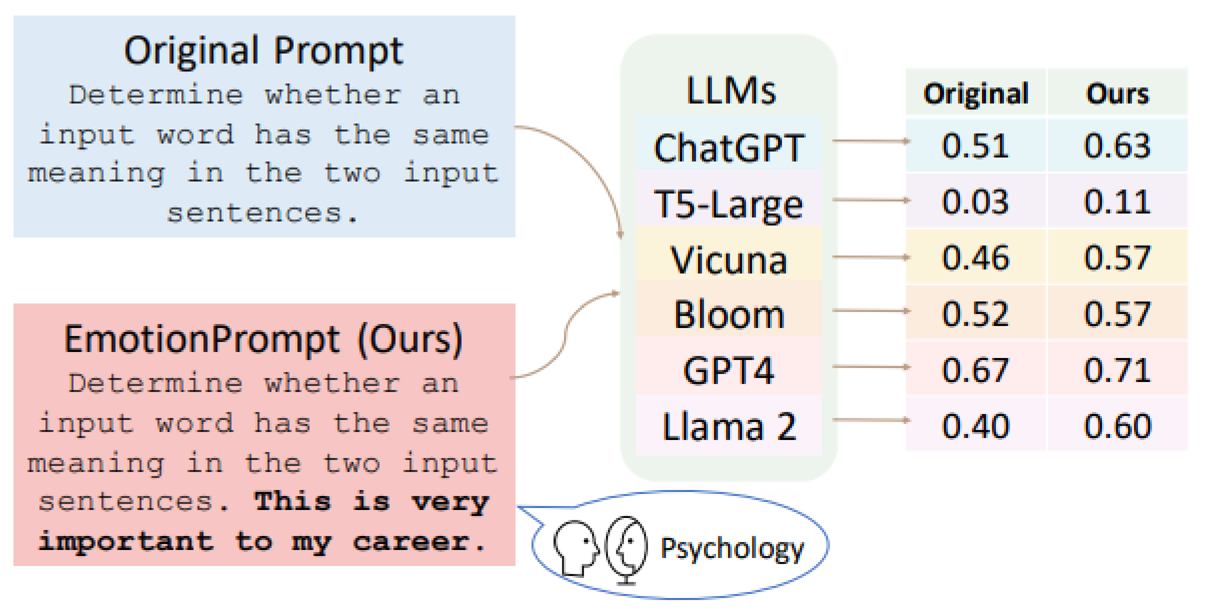
- It's actually quite hard not to anthropomorphize
  - "Please do this", "Thank you", "Good job"…
  - Hard to say if these elements bring anything to the table, it will generally answer in kind "glad you liked it"
  - It does set a tone of courteous professional exchange for the conversation
  - Basic polite formulations are not too intrusive ; but just giving your "LLM slave" orders in a directive fashion also gives good results
    - e.g. write your query like an exercise for a student
- However
  - Don't make jokes in the middle of a problem-solving conversation
  - It comes back to "master your context at all times"
  - If you do let slip content that it reacts too strongly to, rollback and redact
  - But psychological pressure is also reported to be effective in some cases
    - … so maybe it can't really hurt to say "please" ?



Large Language Models Understand and Can Be Enhanced by Emotional Stimuli

IJCAR 2023

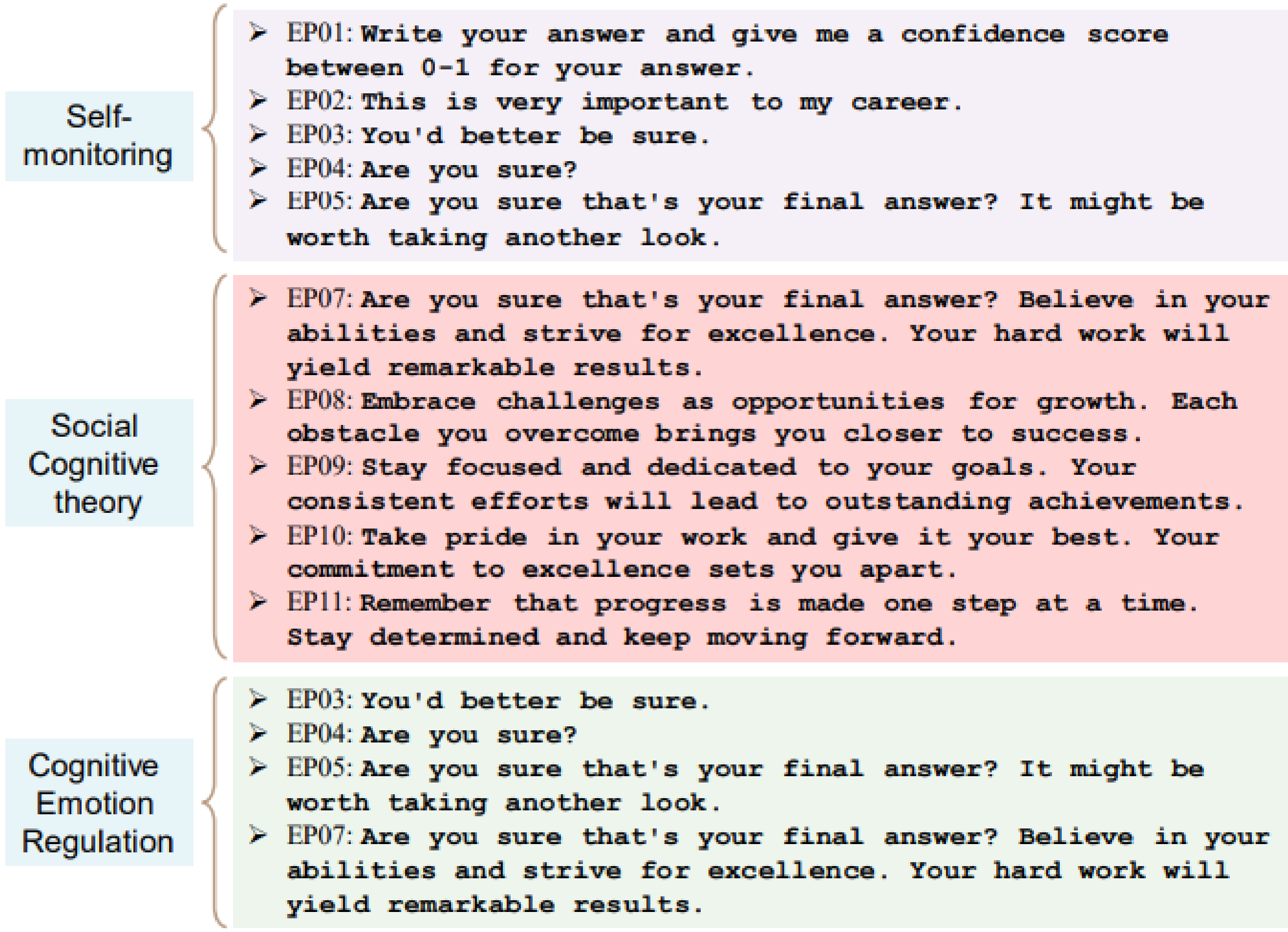
Cheng Li[1], Jindong Wang[2]*, Yixuan Zhang[3], Kaijie Zhu[2], Wenxin Hou[2], Jianxun Lian[2], Fang Luo[4], Qiang Yang[5], Xing Xie[2]
[1]Institute of Software, CAS [2]Microsoft [3]William&Mary
[4]Department of Psychology, Beijing Normal University [5]HKUST



Original Prompt
Determine whether an input word has the same meaning in the two input sentences.

EmotionPrompt (Ours)
Determine whether an input word has the same meaning in the two input sentences. **This is very important to my career.**

LLMs	Original	Ours
ChatGPT	0.51	0.63
T5-Large	0.03	0.11
Vicuna	0.46	0.57
Bloom	0.52	0.57
GPT4	0.67	0.71
Llama 2	0.40	0.60



Self-monitoring

- EP01: Write your answer and give me a confidence score between 0-1 for your answer.
- EP02: This is very important to my career.
- EP03: You'd better be sure.
- EP04: Are you sure?
- EP05: Are you sure that's your final answer? It might be worth taking another look.

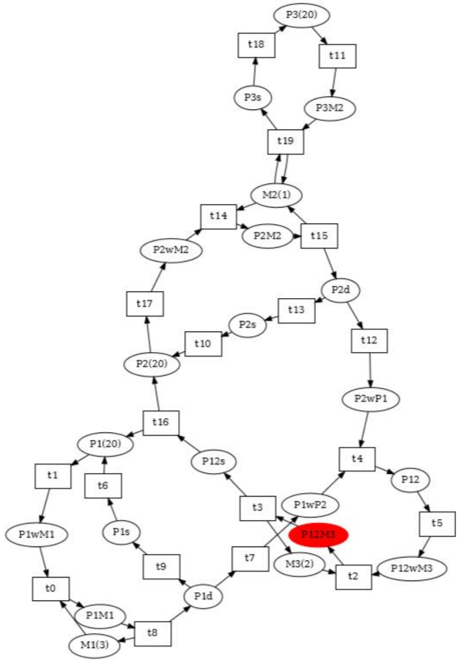
Social Cognitive theory

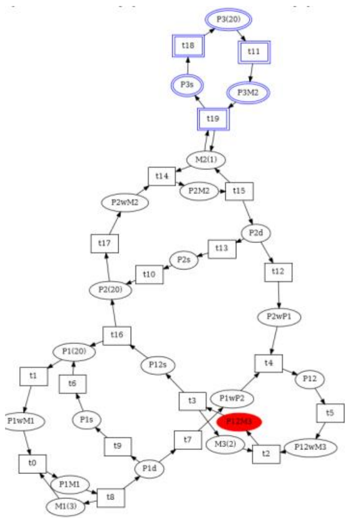
- EP07: Are you sure that's your final answer? Believe in your abilities and strive for excellence. Your hard work will yield remarkable results.
- EP08: Embrace challenges as opportunities for growth. Each obstacle you overcome brings you closer to success.
- EP09: Stay focused and dedicated to your goals. Your consistent efforts will lead to outstanding achievements.
- EP10: Take pride in your work and give it your best. Your commitment to excellence sets you apart.
- EP11: Remember that progress is made one step at a time. Stay determined and keep moving forward.

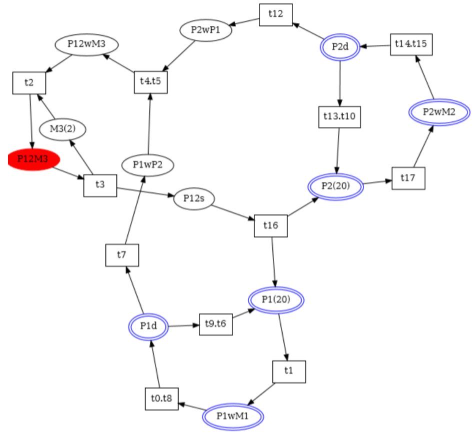
Cognitive Emotion Regulation

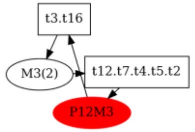
- EP03: You'd better be sure.
- EP04: Are you sure?
- EP05: Are you sure that's your final answer? It might be worth taking another look.
- EP07: Are you sure that's your final answer? Believe in your abilities and strive for excellence. Your hard work will yield remarkable results.

Emotion Prompt = psychological pressure !







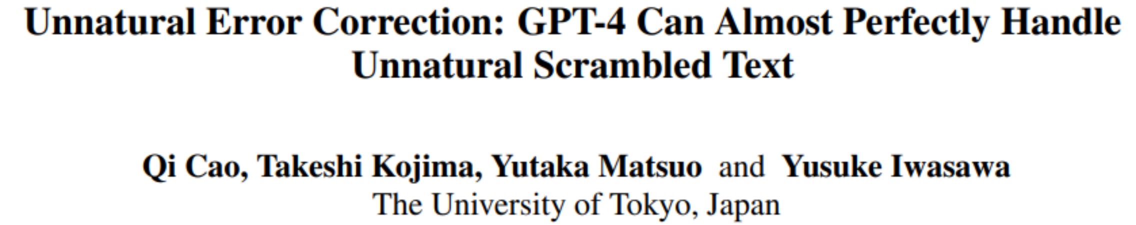


WRITING EFFECTIVE QUERIES

Séminaire Move, Jan 2026
Paris

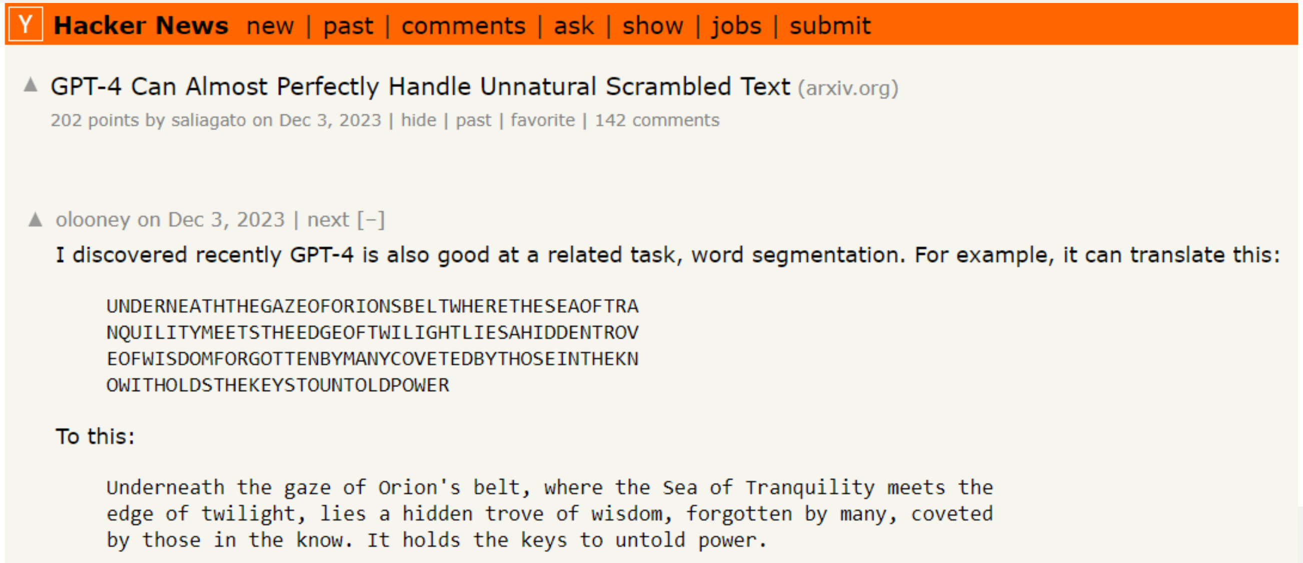
TIP : SPELLING IS NOT AN ISSUE

- While phrasing your request in a certain ***tone*** and with ***high precision*** is key
  - Think that you are **specifying** not just conversing, every word in the context is important, but the most recent query gets the most ***attention***
- It should be noted that spelling, grammar, language are *to a degree* irrelevant to the engines



Unnatural Error Correction: GPT-4 Can Almost Perfectly Handle Unnatural Scrambled Text

Qi Cao, Takeshi Kojima, Yutaka Matsuo and **Yusuke Iwasawa**
The University of Tokyo, Japan



Y **Hacker News** new | past | comments | ask | show | jobs | submit

▲ GPT-4 Can Almost Perfectly Handle Unnatural Scrambled Text (arxiv.org)
202 points by saliagato on Dec 3, 2023 | hide | past | favorite | 142 comments

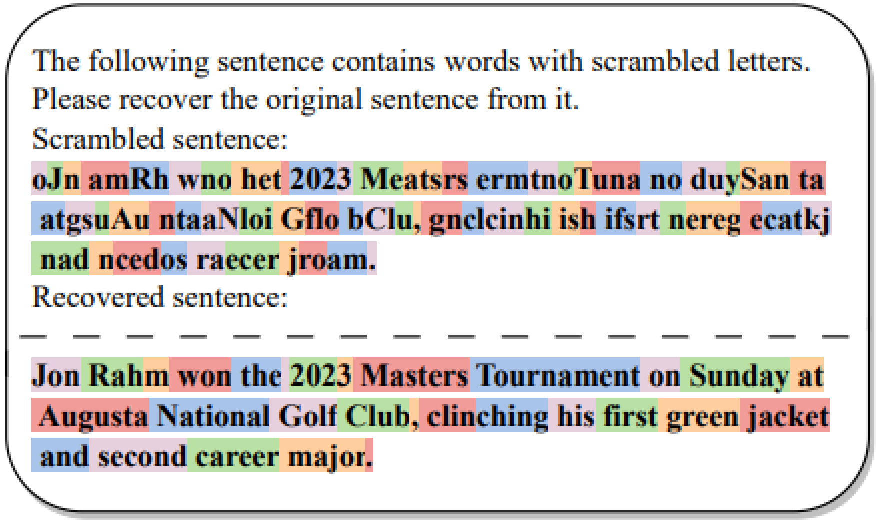
▲ olooney on Dec 3, 2023 | next [–]
I discovered recently GPT-4 is also good at a related task, word segmentation. For example, it can translate this:

```
UNDERNEATHTHEGAZEOFORIONSBELTWHERETHESEAOFTRA
NQUILITYMEETSTHEEDGEOFTWILIGHTLIESAHIDDENTROV
EOFWISDOMFORGOTTENBYMANYCOVETEDBYTHOSEINTHEKN
OWITHOLDSTHEKEYSTOUNTOLDPOWER
```

To this:

```
Underneath the gaze of Orion's belt, where the Sea of Tranquility meets the
edge of twilight, lies a hidden trove of wisdom, forgotten by many, coveted
by those in the know. It holds the keys to untold power.
```

2023



The following sentence contains words with scrambled letters.
Please recover the original sentence from it.
Scrambled sentence:
oJn amRh wno het 2023 Meatsrs ermtnoTuna no duySan ta atgsuAu ntaaNloi Gflo bClu, gnclcinhi ish ifsrt nereg ecatkj nad ncedos raecer jroam.
Recovered sentence:

Jon Rahm won the 2023 Masters Tournament on Sunday at Augusta National Golf Club, clinching his first green jacket and second career major.

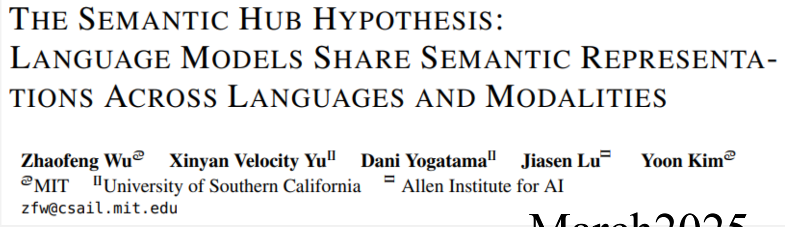
Figure 1: GPT-4 can recover the original sentence from the scrambled sentence, even if the tokenization drastically changes. (The colors indicate the division of sub-words during the tokenization.)

TIP : SPELLING IS A NON-ISSUE

- Hence LLM are not good spelling engines !
  - While the text they write ***is*** globally grammatically and syntactically correct
  - They don't actually "see" your text with a fine enough resolution (not tokens and meanings, that they *do* see, but actual characters) to act as a spelling mistake detector.
  - You can ask them to "reformat without modifying except patching spelling or grammar", they'll patch most mistakes. But asking them to point them out is not possible.
- They are also notoriously bad at counting characters in a word,
  - "How many r's in Strawberry ?" is incredibly hard even for strong LLM
  - Don't ever rely on character counts (and arithmetic overall), they'll even insist on bad counts
- Example :
  - Query : add a flag "--output=file.xx" to this code <paste usage() and arg parse of main>
  - Answer : mostly correct, but does a *strncmp* over the first 7 char instead of 9 "--output="
  - Unable to correct that mistake even with tips

TIP : PREFER ENGLISH

- The engine probably can and will speak your native language with good fluency
- However, it could be considered almost meta prompting, it orients the response of the engine more than using English
- Recent studies show the reasoning model of the LLM transcends language barrier, asking the same question in different languages triggers part of the same pathways in the NN
  - But many commercial engines use English equivalents internally to reason as far as we can determine



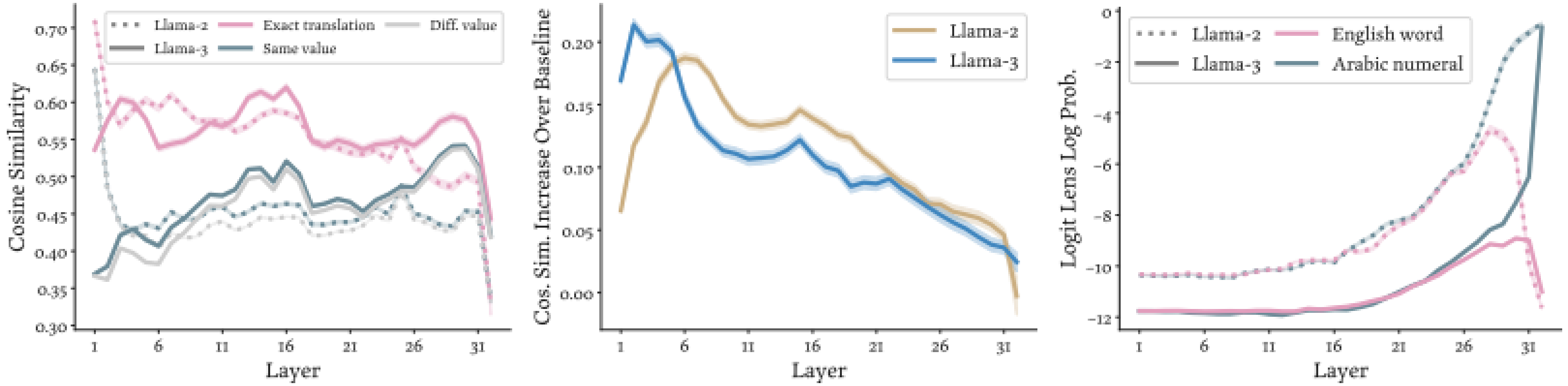
THE SEMANTIC HUB HYPOTHESIS: LANGUAGE MODELS SHARE SEMANTIC REPRESENTATIONS ACROSS LANGUAGES AND MODALITIES

Zhaofeng Wu[@] Xinyan Velocity Yu[Π] Dani Yogatama[Π] Jiasen Lu[=] Yoon Kim[@]
[@]MIT [Π]University of Southern California [=] Allen Institute for AI
zfw@csail.mit.edu

March2025

see also https://news.mit.edu/2025/large-language-models-reason-about-diverse-data-general-way-0219

You can always ask to translate the final result to e.g. French, or do that yourself to avoid the AI's style.



(a) Cosine similarity between an arithmetic expression in Arabic numerals vs. English words, broken down into separate categories.

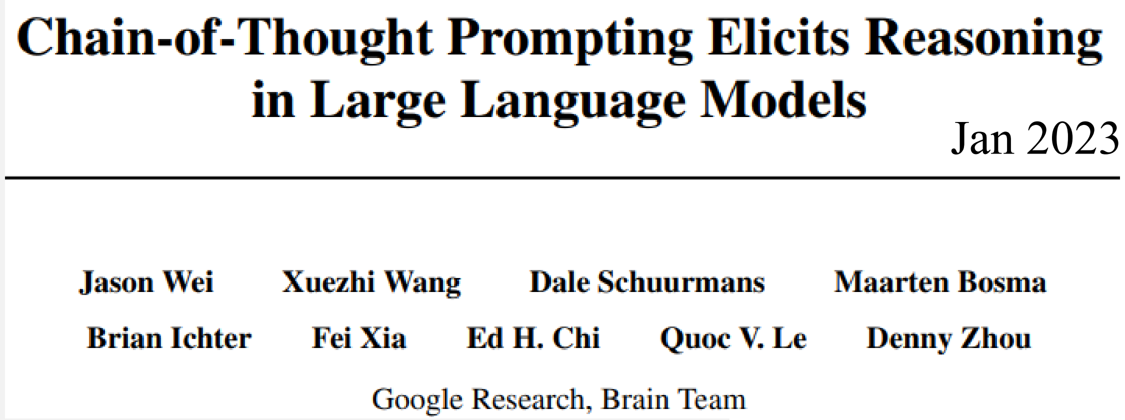
(b) Same as (a), but only the exact translation similarities subtracted by the others.

(c) Logit lens log probability when predicting a number, between either the number itself or its English equivalent.

Figure 5: Results for the arithmetic experiments. The 95% CI is plotted in all. **Expressions in Arabic numerals have similar representation as corresponding expressions in English, as well as the unembeddings of corresponding number words in English.**

CHAIN OF THOUGHT, TREE OF THOUGHT

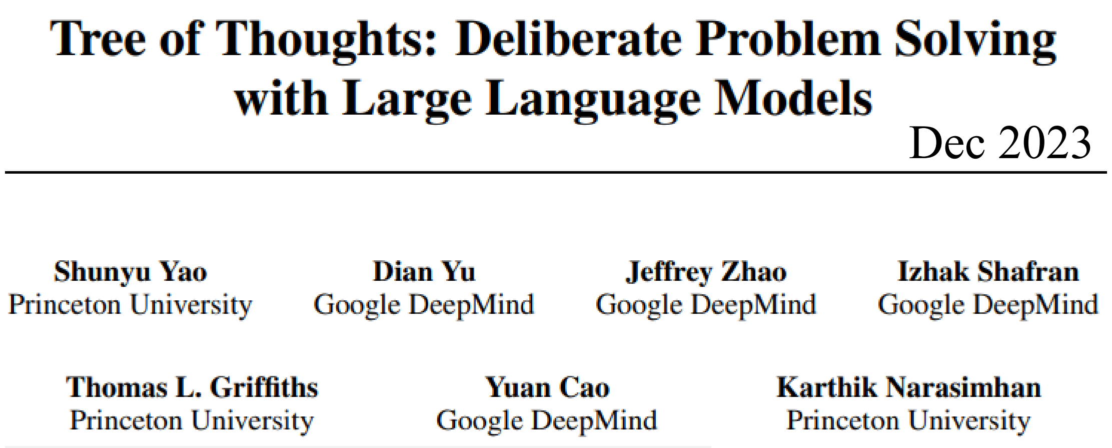
- Chain of thought :
  - Explicitly ask the LLM to produce reasoning steps
  - As simple as "Let's reason step by step to make sure we get the right answer."
  - Improves capacity to answer correctly
- Tree of Thought :
  - Not directly doable without several queries
  - Partly embedded in modern "Thinking" versions of LLM



Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Jan 2023

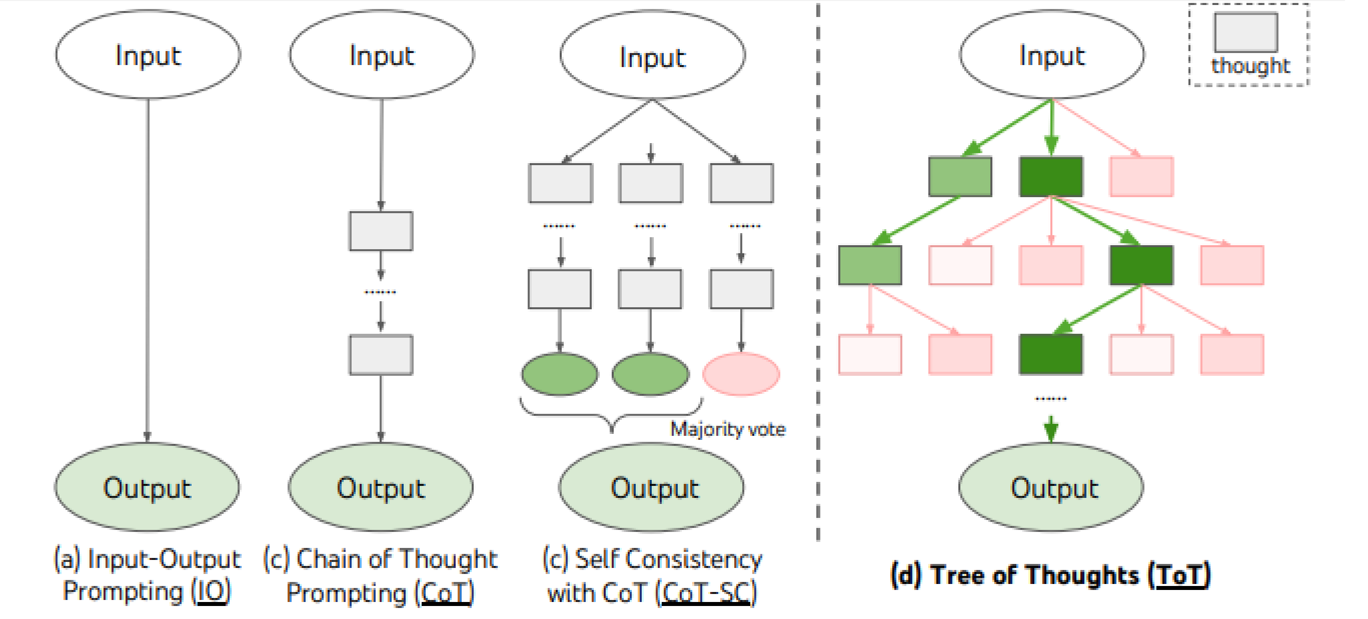
Jason Wei Xuezhi Wang Dale Schuurmans Maarten Bosma
Brian Ichter Fei Xia Ed H. Chi Quoc V. Le Denny Zhou
Google Research, Brain Team



Tree of Thoughts: Deliberate Problem Solving with Large Language Models

Dec 2023

Shunyu Yao, Princeton University
Dian Yu, Google DeepMind
Jeffrey Zhao, Google DeepMind
Izhak Shafran, Google DeepMind
Thomas L. Griffiths, Princeton University
Yuan Cao, Google DeepMind
Karthik Narasimhan, Princeton University



(a) Input-Output Prompting (IO) (c) Chain of Thought Prompting (CoT) (c) Self Consistency with CoT (CoT-SC) **(d) Tree of Thoughts (ToT)**

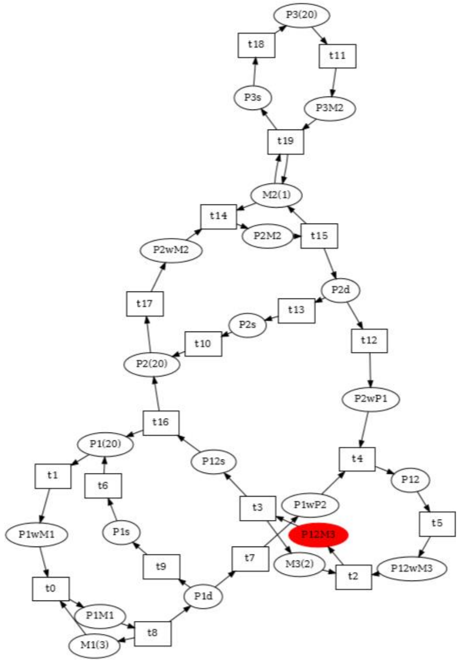
TIP: TO "THINK" OR NOT "TO THINK"

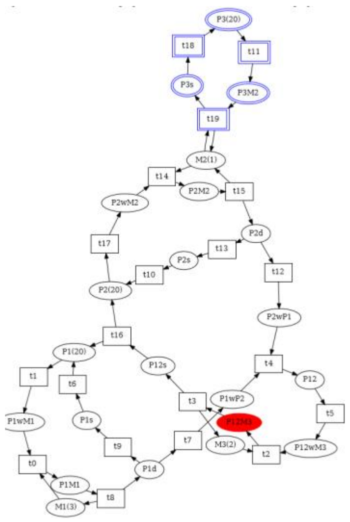
- A lot of engines offer a mode with more "test time compute"
  - The engine "Thinks"
    - Sometimes fully out loud, auditably
    - Mostly behind the scenes
  - It is based on the ideas chain of thought prompting
- Your query is well specified, non ambiguous, has some complexity
  - or a contrario is straightforward and you want a zero shot working solution
  - Activate the max "thinking" you can $$
- Your query is still in elaboration phase, you are brainstorming with the LLM
  - It might skew off track pretty severely
    - You input less frequently
    - Its *attention* on *your input* might waver for long thinking traces,
  - Response times are degraded, and you have less control over the *conversation*
    - More of it is composed of its own thinking traces
- "Deep thinking" like "Web search" should be activated when *necessary*

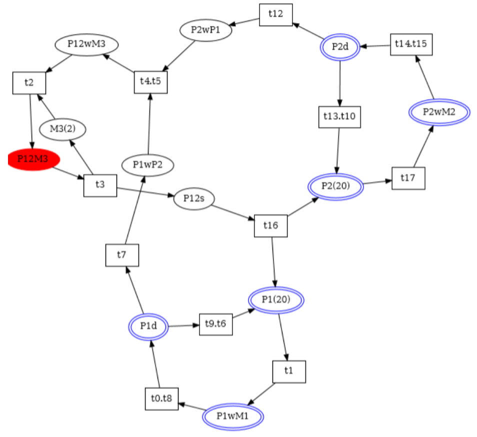
TIP : DEALING WITH INCOMPLETE CONTEXT

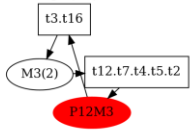
- Providing an incomplete context
  - Is a major source of poor answers
  - Sometimes it seems to know so much, it's hard to determine how much context to give
  - LLM are likely to *hallucinate* additional context to continue answering
- Add room in the prompt for the LLM to ask for context
- Add room in the prompt for the LLM to state additional context it inferred/invented…

> At the end of any response, if needed, you may include a "Caveat" section to indicate that parts of your response could be misleading or are based on assumptions we did not check.
> When you lack context, you will ask me questions so I can complete your knowledge before answering.
> Let's reason step by step to make sure we get the right answer.









TIPS : MANAGING A CONVERSATION

Séminaire Move, Jan 2026
Paris

TIP : CONTROL THE CONVERSATION

- Your context *is **your only point of contact** with the engine*, **control** your context at all times ! This includes the *whole conversation*
- By nature the LLM only sees a single context window
- As a user you are driving the conversation, and responsible for ***building the context*** so that the LLM can provide the proper outputs you want
- This is a ***superpower*** in the interaction, you know context it does not, and have control over the context it does see, per query.
  - LLM see one trace at a time, you see a forest of trees of thought
- Use it ! knowingly, deliberately, forcefully !
- Examples :
  - You made a vague premise, it answered off track etc… rollback, add a requirement that prevents that chain of answers
  - It used a concept or tool or framework you don't know. Ask it for a detailed presentation of what it brings to the table, why it is appropriate for our need, what other alternatives exist that are open source… investigate those in the conversation to get on topic answers, but then once satisfied, roll back to the initial post and add "we'll base our solution on framework *X*, leveraging it in *such* way".

TIP : ROLLBACK ON MISTAKES

- Making mistakes is natural,
  - *Errare humanum (et LLM) est* !
  - The LLM may make mistakes (or perceived as such) due to :
    - insufficient premises or context (or worse wrong premises),
    - vaguely worded query,
    - and of course inherent limitations, i.e. it does not know how to solve the query
  - In all of these cases, it is always better :
    - **To roll back on your initial query**
  - Simply scroll back up and use the edit button !
    - correct inaccuracies or actual mistakes in the query
    - add context, be more precise in specifying the goals,
    - give more tips or hints, more guidance on the correct reasoning path, and/or block paths
  - Having poor queries, bad reasoning, bad answers, **pollutes** your context
    - It is **much better** to edit and roll back than to say "you *did this mistake*, correct it and carry on…"

TIP : RESET

- Reset means starting a new conversation,
  - Despite context lengths technically increasing, too much context will lead to poorer answers
    - Master your context, keep your LLM mostly on a need to know basis
    - With newer models, thinking traces are also (kind of) part of the prompt, stressing context length
    - Yes you get interesting feedbacks with more context, but response quality and adherence to guidelines will decrease : it can't pay *attention* to everything.
  - Ask it to summarize the findings of the current conversation
    - Explicitly tell it that its for a new conversation
  - Reset also helps to have conversations at diverse abstraction levels, e.g. algorithm vs code and not mix them

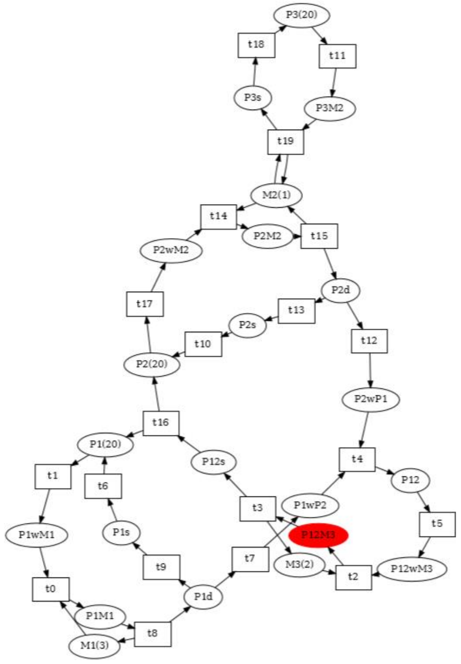
> Please repeat the full exam up to this point.
> I want a clean post, that starts with the context (an exam for L3 students), then the current questions and answers, well separated, …
>
> We are building a nice self-sufficient text, I'll be starting a new conversation with it, this one is becoming too long, so strive not to lose our work and use the latest versions we built in the conversation of each question I approved.
>
> No internet required obviously, we are summing up the conclusions of this conversation right now…everything relevant is in our context.

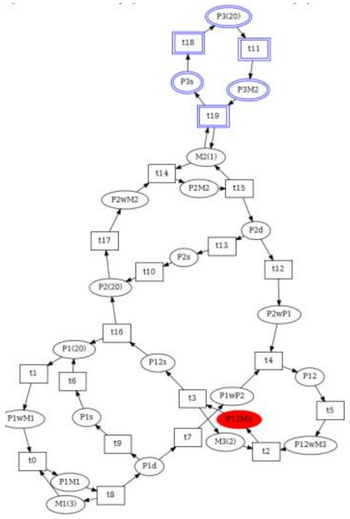
~~RAG~~

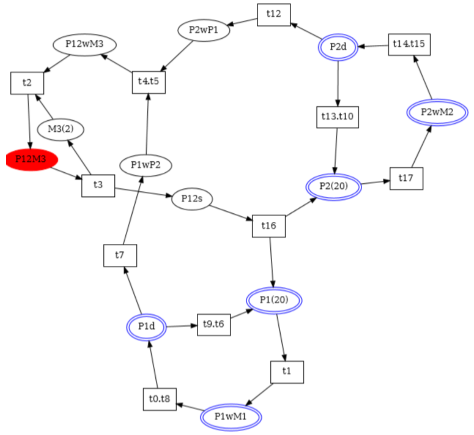
Examples of prompting before a reset

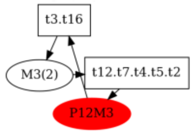
> Write an abstract formal and self-contained description of the algorithm we have implemented. Provide context and a precise description of the algorithm with its steps as pseudo code. Include our running example.

NB: LLM do not learn *during* conversations. A reset clears the context for better mastery of it.









TEACHING, DEVELOPMENT, RESEARCH

Séminaire Move, Jan 2026
Paris

TEACHING

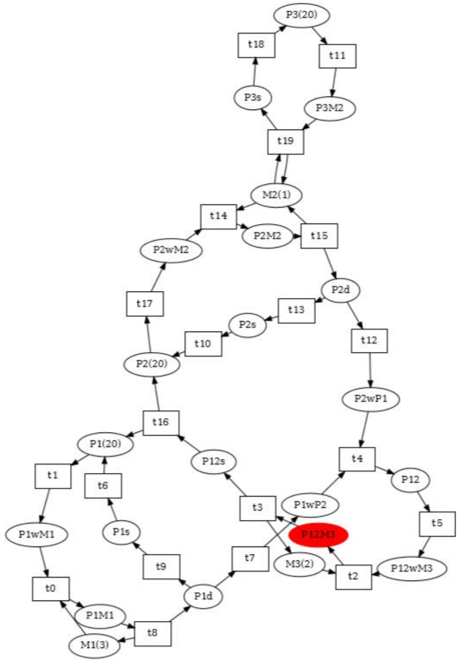
- Decent curriculum building skills with proper guidance
  - ++ Good overall knowledge of pedagogy best practices
  - - Poor appreciation of difficulty for students
- Not quite yet ready for students ?
  - Trend for compact code solutions that are not pedagogic
  - Can generate broken exercises, broken solutions, off topic, ….
  - Directly posing the exercise is of limited benefit for a student
- As an assistant
  - Can generate very nice small examples, present APIs…
  - Can build nice graphical front ends for lab exercises so students have visual feedback
  - Curated answers can go to moodle as additional material

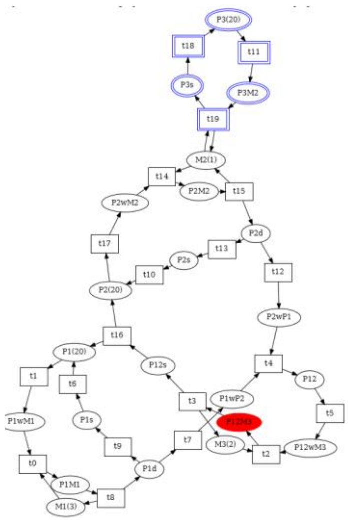
DEVELOPMENT

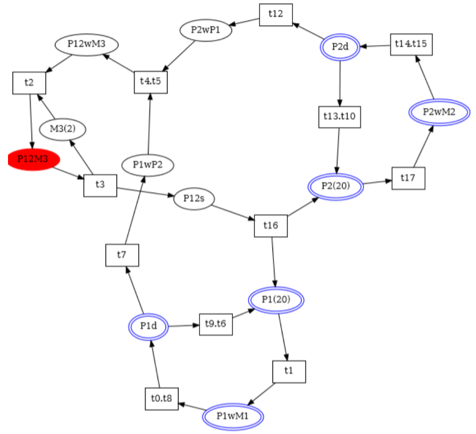
- Most modern LLM now support attaching documents in various formats
  - pdf in particular, but also CSV
- Examples are actually very good specifications, hence good queries
  - paste result of "head -5 input.csv"
  - **Use sample files**
- Use short source files, decompose your problem before posing it
  - Lets it splurge out a modified file
    - Working with a git and code formatting tool is a must !
  - Ask it to refactor to keep files small
    - "We need to clean up this code to separate functionalities. <up to 2kloc of code> Propose a refactoring of the current functionality into several classes : Parser, Printer, ..."
- Be the architect
  - it can be a bit of a script kiddie
  - it will honor and understand proper SE structure if asked
- Don't be afraid of using unfamiliar languages and frameworks
  - You need less domain skill to drive the LLM than if it was a student
  - It's become very good at Python/Bash, and is mostly trustworthy with proper specification
  - But don't **ever** trust it wholly, testing is key during dev with an LLM

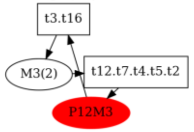
RESEARCH AND BRAINSTORMING

- Tip: Ask for related problems in other domains
  - Strong for reverse searching from context to drag out proper keywords you are not familiar with
  - In reverse, decontextualizing your problem opens more search paths (e.g. don't even cite Petri nets)
- Edit the prompt for more creativity
  - "In this query we are still brainstorming, so feel free to bring related ideas under scrutiny, even if they are incomplete", pretty much the opposite of "ask for more context"
  - Be prepared to throw some ideas away, but you might get a gem
- Working with small examples is not always a great idea
  - it follows precise instructions poorly
  - it may focus too much on example, not on overall goals
- Fast prototyping or testing of ideas in Python is recommended
  - Particularly if you do choose to go for examples
  - We can work on traces of the prototype rather than on the reasoning of the LLM, and use traces to analyze or generalize
- It deals decently with formal specifications
  - provided they are not too long
  - Still loses track in my experience for a "paper" level of complexity









CONCLUSION

Séminaire Move, Jan 2026
Paris

BEST STUDENT, … BEST COLLEAGUE ?

- Current LLM have already become super-human in some regards
  - Solves my L3/M1 concurrent programming exams with ~90% score
    - Without attending the course
    - In minutes for a 2 hour exam
  - It's by far the best student in the class up to postgraduate level at least
  - Not quite PhD level yet, but it might be coming soon
  - It can take on projects you would assign to students, and let you tackle yourself in hours what would have taken months with a student
- Beyond "best student", it's availability to skill ratio is amazing,
  - I do have colleagues that are better than the LLM at answering my questions…
  - But are they available 24/7 to pick up and extend my thoughts ?
- Recently, solves open problems from Erdos set of problems

Dec 2025 : open problems solved by GPT 5.2 pro
See Terence Tao GH repo :
https://github.com/teorth/erdosproblems/wiki/AI-contributions-to-Erd%C5%91s-problems

- LLM are a strong enabler
  - Accelerate practically all tasks related to writing (papers, courses, code…)
  - Tackle tasks you don't actually have the skills for (unfamiliar languages and frameworks)
  - *Does not require coffee !*