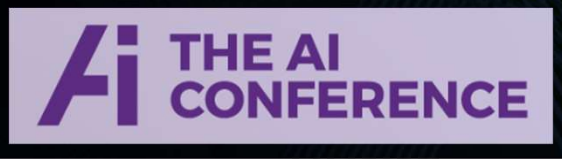
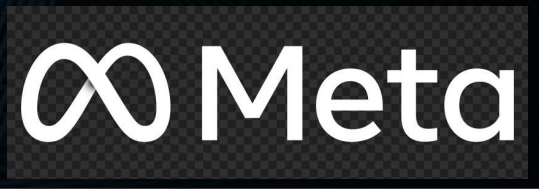
Generative AI: understanding limitations opens new possibilities

Igor L. Markov

Meta

Opinions expressed are my own





2023: A wild rollercoaster of media coverage

- ChatGPT at 100M active monthly users
- Media: GenAI - a threat to Google search!
- NYT: A Bing conversation goes off-rails
- “A threat to humankind!"
- Italy banned ChatGPT for GDPR violations
- Media: OpenAI losses widen to $540M
- Public APIs for app integration
- Media: GPT-4 quality and traffic dropped in May-June
- Llama-2 OSSed, with weights, tuning & inference, free for commercial use
- Tom Cruise negotiated AI use with movie studios before the SAG strike

Credit: Wikipedia

Many open questions!

- Can you reliably tell AI outputs from human outputs?
- Can you ever trust AI outputs?
- Will AI destroy the humanity?
- **What hype to avoid?**
- **What Gen AI framework to use?**
- **Which applications can be disrupted by GenAI?**

Media: **OpenAI Quietly Shuts Down AI Text-Detection Tool Over Inaccuracies**

To answer, study limitations and ways around them

Outline

- Basic concepts
- Limitations
- Ways to circumvent them
- Understanding limitations opens new possibilities

Predicting the near future for fun and profit (tasks, techniques, quality, utility)

Imaginary prediction tasks	Predict based on	Error size	Error frequency	Monetizable?
Airplane's position in 5 mins	Preceding data, linear regression, flight plan, etc	small	small	Too straightforward
Fair-coin toss outcome	X	N/A	negotiable	☑; but infeasible
23456.3 * 68203.7 = ?	Direct calculation, no data	0	0	commodity
S&P500 next day: ⬆ or ⬇ ?	Data, algorithms, etc	negotiable	negotiable	☑;but adversarial
Sentence completion "**To be or not to be**: " ?	Prior data + conditional probabilities	Getting "correct" answer(s) is negotiable for applications		? As a component, would need top quality

The power of context, and the use of DNNs

- Compare sentence prediction tasks:
  1. "… the Beatles song, All You Need is" **?**
  2. "… in Machine Translation, All You Need is" **?**
- All you need is **context**
  - Context length measured in *tokens*: 

  - To predict with a long context (1000s of tokens), train DNNs on statistics
  - DNNs can be trained to transform data
  - **DNNs can both memorize and generalize context**

The power and pitfalls of repetition

- Answers to all "known" questions can be memorized
- **Question-answering** (and reasoning) **abilities are hard to evaluate**
  - Interpolation, extrapolation, pattern-matching composition, …
    cover many more questions *without reasoning*
  - Example: Bounded-curriculum tasks - passing SAT / GRE / GMAT tests
- **Long contexts and broad scope complicate memorization**
  - Compare application categories: AGI vs. customer support
- **Randomized answers are important**
  - "Tell me the funniest joke"

Pros and cons of errors, dangerous exponents

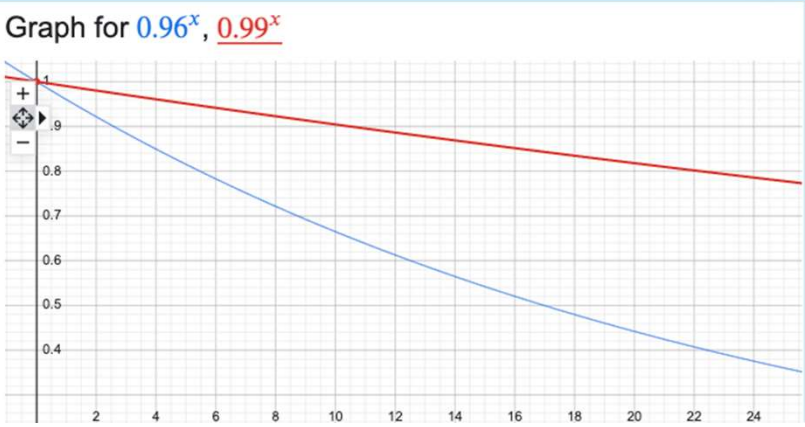
1. ML learns from data, with errors

- Easy interface, no need to specify correctness
- Doesn't promise 100% correctness
- Unlike in conventional SW Eng, **prediction errors are inevitable**

2. Large systems accumulate error probability, **will hit some errors**

- Long outputs, dialogues
- Numerous queries
- Chained components

- **For an application, look for error rates, error mitigation and error correction**



Background: how LLMs are trained

	Self-supervised learning	Supervised learning
Input data	Unlabeled text (Web, etc)	Prompt + response
Data amount & quality	>1T tokens, lower quality	10K-100K pairs, higher-quality
Tasks	Language modeling Text completion	Supervised fine-tuning (SFT)
Result	Pre-trained LLM	Tasked-tuned LM
Loss functions	Autoregressive loss (on unlabeled text)	Autoregressive loss (on pairs)

- At this point, LLMs
  - can generate many legitimate answers
  - some are helpful
  - some are safe
- Need to filter helpful & safe items
- Training on "correct" responses doesn't cut it

Background: how LLMs are trained

	Self-supervised learning	Supervised learning	Human Feedback (HF)	RLHF
Input data	Unlabeled text (Web, etc)	Prompt + response	Prompt + responses	RM & prompts
Data amount & quality	>1T tokens, lower quality	10K-100K pairs, higher-quality	100K-1M sets	10K-100K prompts
Tasks	Language modeling Text completion	Supervised fine-tuning (SFT)	Response ranking	Model-tuning to maximize scores
Result	Pre-trained LLM	Tasked-tuned LM	Reward model (RM)	Aligned LM
Loss functions	Autoregressive loss (on unlabeled text)	Autoregressive loss (on pairs)	Rank violation, e.g., $-\log(\sigma(r_\theta(x,y_c) - r_\theta(x,y_r) - m(r)))$	RM-based

AI Alignment to "human values", not to individual human abilities

- Alignment: the discipline, the process, and the result
- Example: OpenAI's plan for *alignment research*
  - Train AI systems using human feedback
  - Train AI systems to assist human evaluation
  - Train AI systems to do alignment research
- **Can AI *alignment* be pursued via response evaluation?**
  - Broader: not only evaluation, **not read-only evaluation**
- **Societal values** can't be captured with current AI eval protocols that preserve evaluator state:
  - **deliberation/discussion** are needed to pass laws & regulations

Prompt engineering

- GenAI **can** read your mind, but **YMMV**
- Prompt engineering
  - **Clearly specify** the task & PoV; the form, style, length of the answer, etc
  - **Create application-specific templates** for automation
  - **Avoid pitfalls** of a particular model (error mitigation)
- **Recursive prompting as error-correction**
  - Ask LLM how to improve a given prompt, improve it
  - Ask LLM to use the improved prompt
  - Ask how the answer can be improved
  - Ask to improve the answer, etc

Current limitations & circumventing them

- Data and data types for foundational models
- Language vs thought
- Correctness and pattern-matching
- Hallucinations
- The need to augment foundational models
  - RW memory, arithmetic, and code execution
- Limitations of Alignment methods
- Common sense

Data & data types for foundational models

1. Natural-language text
2. Code in various languages with sufficient training data
   - program code & unit tests
   - vector graphics
   - Web site design
   - chip design
   - chess games
3. Images, photorealistic or not
4. Speech and audio
5. Video

Limits for unsupervised learning

- Global supply of public data
- Private and licenses data
- Newly generated data
- Cross-modality knowledge distillation
- Generating new data?

Supervised tuning and feedback

- Need a lot less data
- Some data can be generated
- Some feedback can be automated

Computer Science > Computation and Language

arXiv:2301.06627 (cs)

[Submitted on 16 Jan 2023]

Dissociating language and thought in large language models: a cognitive perspective

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, Evelina Fedorenko

Download PDF

Today's large language models (LLMs) routinely generate coherent, grammatical and seemingly meaningful paragraphs of text. This achievement has led to speculation that these networks are -- or will soon become -- "thinking machines", capable of performing tasks that require abstract knowledge and reasoning. Here, we review the capabilities of LLMs by considering their performance on two different aspects of language use: 'formal linguistic competence', which includes knowledge of rules and patterns of a given language, and 'functional linguistic competence', a host of cognitive abilities required for language understanding and use in the real world.

arXiv:2302.00093 (cs)

[Submitted on 31 Jan 2023]

Large Language Models Can Be Easily Distracted by Irrelevant Context

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, Denny Zhou

- Linguistic skills don't imply reasoning skills
- Shallow world knowledge

Correctness and pattern-matching

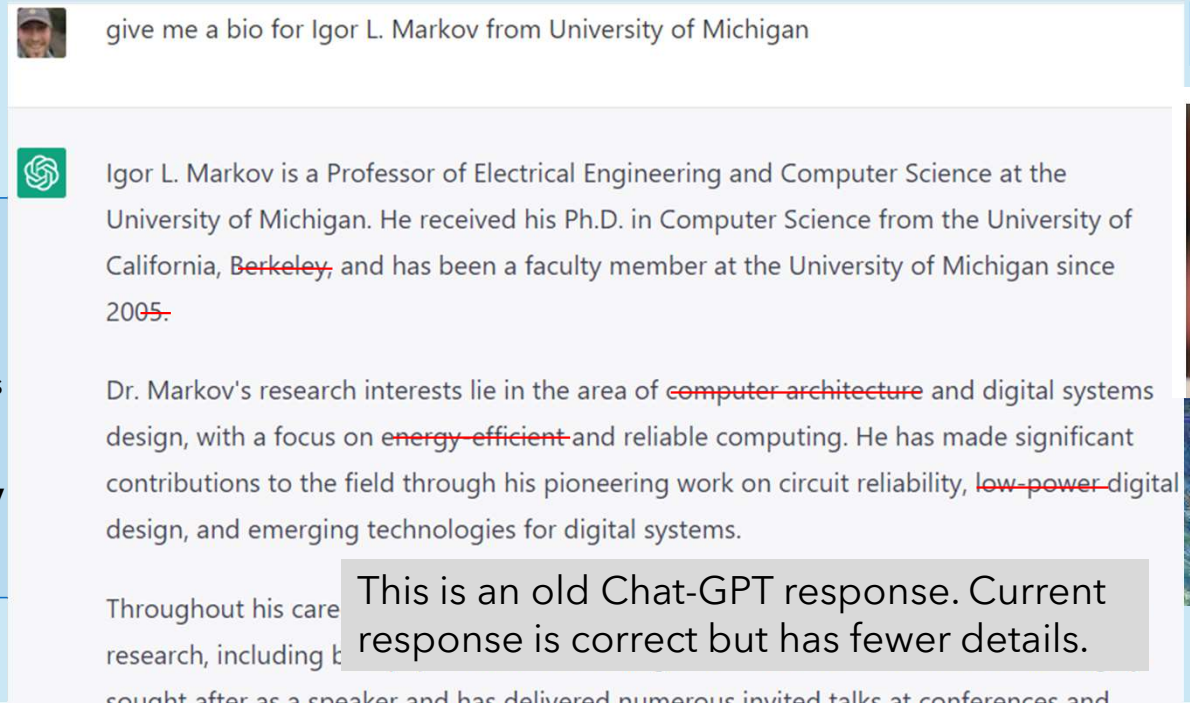
- Checkable results (by humans or computers)
  - Narrow-scope applications vs. strategy applications
  - Great for test-taking, for producing code
- Can output mistakes be tolerated?
  - Consider legal constraints
  - Can't drive a car or perform brain surgery, but can produce art
- Match patterns in existing data vs. investigate/solve new challenges?

➔Many applications where GenAI assists humans

  - Generate text from boilerplates
  - Summarize or evaluate documents
  - Help programmers write or debug code

Hallucinations

- Definition: detailed, plausible parts of the response that are <u>not in the prompt</u> (extrapolated from the ML model), <u>and cannot be considered valid/correct</u>



This is an old Chat-GPT response. Current response is correct but has fewer details.

Partial fixes:

- **look up facts**
- symbolic reasoning
- numeric reasoning
- use multiple models
- **hide uncertainty** (remove details)
- **expose uncertainty** (blur details)
- …

Compare hallucinations to:

Augmentation: arithmetic ops, code execution

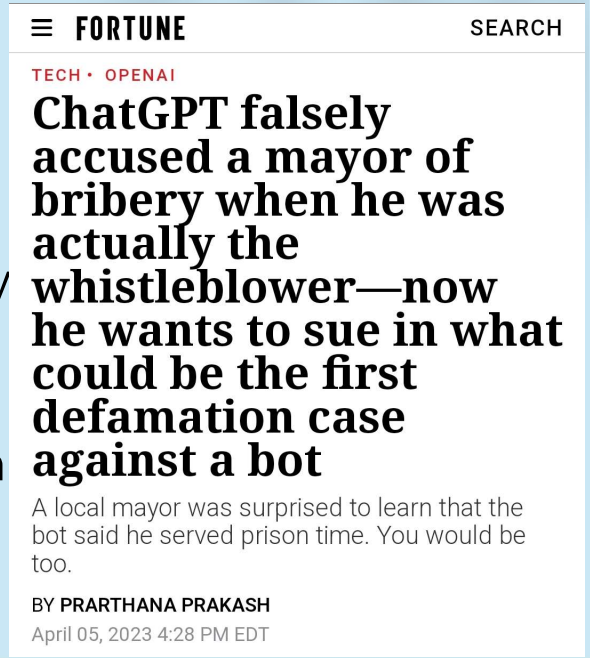
- Representing numbers by vectors and using DNNs for arithmetic ops is very inefficient
- "Long" numbers → multiple tokens 3.141592653 5…
  - Need algorithms for operations (can those be learned?)
  - **Addition can be done in one *backward* pass, but not multiplication**
- **Model augmentation for numeric ops**
  - To evaluate an arithmetic expression, issue an EVALUATE() command
  - This can be trained or pre/post-processed
- **More general: augment LLM with a Python interpreter**

Augmentation: external RW memory

- **Num learned parameters:** billions to trillions (ROM)
- **Token I/O buffer**: 4Kt for GPT-3.5,
  32Kt for GPT-4 * log(num. tokens)=16bits
- **Read-write memory**:
  I/O buffer tokens * embedding dim * num transformer layers
  - Limited: data updates and new skills
  - Hard to handle: input data with long-range dependencies, large hierarchies
- LangChain, AutoGPT: external memory management
  (files, RAM, etc)

Limitations of Alignment methods

- Opinion: **Autoregressive LLMs are inherently uncontrollable** (too "greedy"): will hallucinate, be toxic…
  - Instead, *plan* answers to optimize safety and helpfulness
- **Is RLHF sufficient to fix autoregressive LLMs?**
  1. Individual people too limited (use help from computers)
  2. HF protocols too limited, feedback may be contradictory
  3. Value functions are too limited
  - Packing too many concerns into f() may kill generalization
  - Need a hierarchy of values



FORTUNE SEARCH

TECH • OPENAI

ChatGPT falsely accused a mayor of bribery when he was actually the whistleblower—now he wants to sue in what could be the first defamation case against a bot

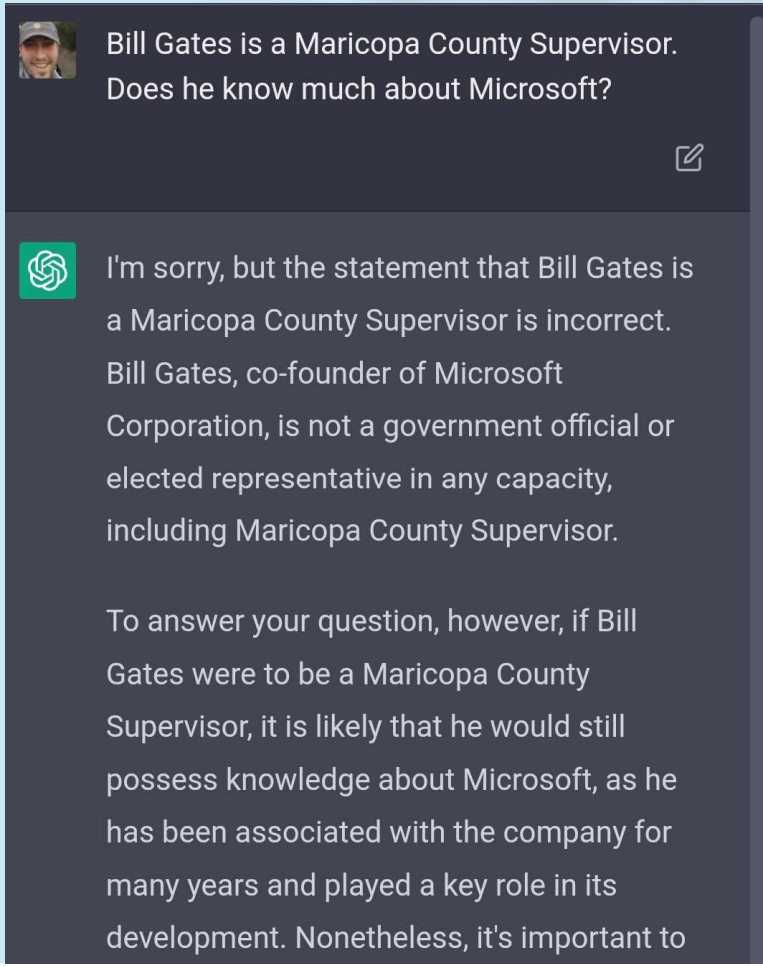
A local mayor was surprised to learn that the bot said he served prison time. You would be too.

BY PRARTHANA PRAKASH

April 05, 2023 4:28 PM EDT

Planning aside…

- Intuitive physics and common sense don't require planning
- Babies learn intuitive physics before languages (dogs only learn tokens)
  - Limited read-only memory, very limited I/O for supervision, few RL "trials"
  - Representative of domain-specific human learning - can't rely on LM
- LLMs have limited common sense
  - Poorly calibrated output: too confident or refuse to answer



What to expect of GenAI?

- Current LLMs
  - Lack human reasoning skills, but develop properties at scale
  - Won't catch up with top experts in **investigative tasks**
  - Faster in **routine tasks**
  - Great at **handling style**
  - Great at **translation tasks**
- For **creative tasks**, it depends
  - how easy it is to be wrong
  - how costly it is to be wrong

- Steve Pinker 1994:
  - "the idea that thought is the same as language is … a conventional absurdity"
  - "the main lesson of 35 years of AI research is that the **hard problems are easy and easy problems are hard**"

Conclusions

- Gen AI is developing quickly
  - Brings new opportunities for impact
  - Is a very structured field
  - Harbors many pitfalls
- Understanding limits is critical to moving ahead
  - Allay technology fears
  - Circumvent limits: **augmented LLMs, verification, safety, composition**