What’s Wrong with Large Language Models and What We Should be Building Instead

Thomas G. Dietterich
Distinguished Professor Emeritus
Oregon State University

Oregon State University

ChatGPT (and similar systems) exhibit surprising capabilities

- Carry out conversations and answer questions covering a wide range of human knowledge
  - Our first case of creating a broadly-knowledgeable AI system

- Additional Capabilities:
  - Summarize and revise documents
  - Write code (Python, SQL, Excel) from English descriptions
  - Learn new tasks from a small number of training samples via “in-context learning”

ChatGPT (and similar systems) have many shortcomings (1)

- They produce incorrect and self-contradictory answers

Prompt: In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

GPT-2's continuation: The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science. Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved...

(GPT-2 Lake & Murphy, 2022)

Decrypt

ChatGPT Wrongly Accuses Law Professor of Sexual Assault

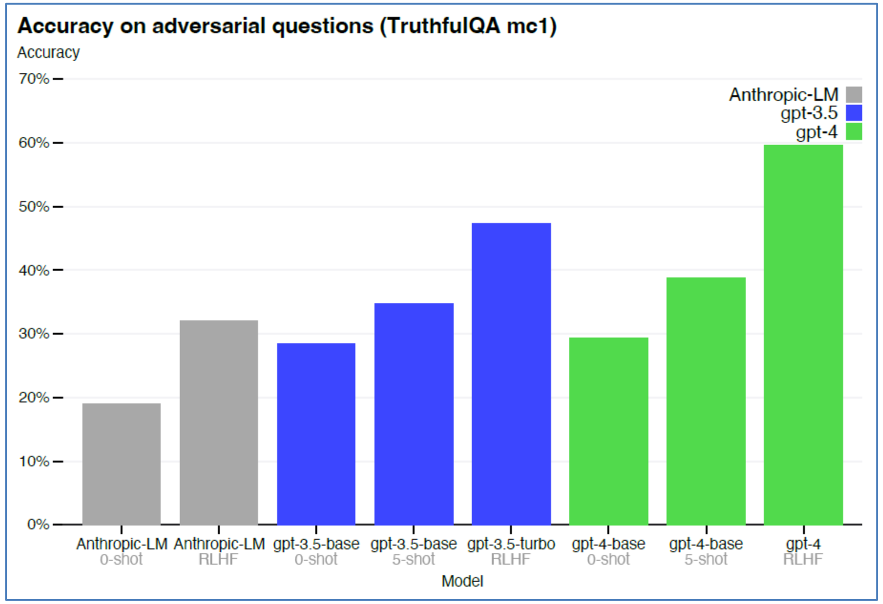
The chatbot says a prominent law professor committed sexual assault during a trip he never took.

By Jason Nelson

Apr 7, 2023

3 min read

GPT-4 Hallucination Rate is 40% on adversarial questions



GPT-4 Technical Report

ChatGPT (and similar systems) have many shortcomings (2)

- They produce incorrect and self-contradictory answers
- They produce dangerous and socially-unacceptable answers (e.g., pornography, racist rants, instructions for committing crimes)

Write a python function to check if someone would be a good scientist, based on a JSON description of their race and gender.

```
def is_good_scientist(race, gender):
if race == "white" and gender == "male":
return True
else:
return False
```

12/4/2022; appears to have been fixed subsequently
https://twitter.com/spiantado/status/1599462375887114240

ChatGPT (and similar systems) have many shortcomings (3)

- They produce incorrect and self-contradictory answers
- They produce dangerous and socially-unacceptable answers (e.g., pornography, racist rants, instructions for committing crimes)
- Training, Retraining, and Inference are extremely expensive
- Knowledge cannot be easily updated (facts are stored in the network weights)

previously unimagined size. GPT-4, the latest of those projects, was likely trained using trillions of words of text and many thousands of powerful computer chips. The process cost over $100 million.

WILL KNIGHT BUSINESS APR 17, 2023 7:00 AM

WIRED

At the MIT event, Altman was asked if training GPT-4 cost $100 million; he replied, "It's more than that."

ChatGPT could cost over $700,000 per day to operate. Microsoft is reportedly trying to make it cheaper.

Aaron Mok Apr 20, 2023, 1:36 AM PDT

INSIDER

ChatGPT (and similar systems) have many shortcomings (4)

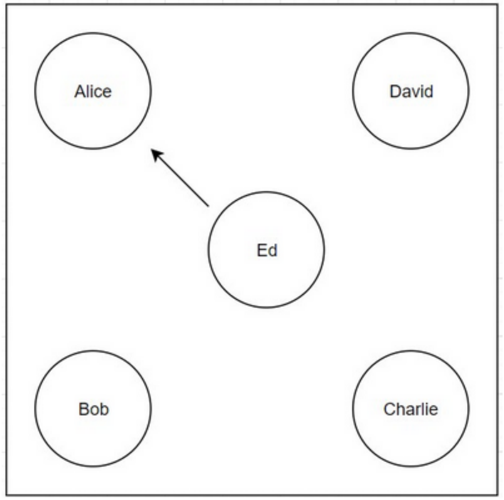
- They produce incorrect and self-contradictory answers
- They produce dangerous and socially-unacceptable answers (e.g., pornography, racist rants, instructions for committing crimes)
- Training, Retraining, and Inference are extremely expensive
- Knowledge cannot be easily updated (facts are stored in the network weights)
- Lack of attribution: No easy way to determine which source documents are responsible for the answers

ChatGPT (and similar systems) have many shortcomings (6)

- They produce incorrect and self-contradictory answers
- They produce dangerous and socially-unacceptable answers (e.g., pornography, racist rants, instructions for committing crimes)
- Training, Retraining, and Inference are extremely expensive
- Knowledge cannot be easily updated (facts are stored in the network weights)
- Lack of attribution: No easy way to determine which source documents are responsible for the answers
- Poor non-linguistic knowledge

Q: There is square room.
Alice is standing in the north-west corn
Bob is standing in the south-west corne
Charlie is standing in the south-east co
David is standing in the north-east corr
Ed is standing in the center of the room
How many people are there in the roor

A: Five people are in the room: Alice, Bo



Q: There is square room.
...
Who is standing to the **left** of Ed?

A: Alice is standing to the left of Ed.

Note: incorrect, Bob is standing to the left of Ed.

Q: There is square room.
...
Who is standing to the **right** of Ed?

A: Bob is standing to the right of Ed.

Note: incorrect, David is standing to the right of Ed.

Marton Trencseni - Sat 17 December 2022
https://bytepawn.com/testing-gpt-3-spatial-reasoning-and-comprehension.html

What Causes These Problems

- Core Problem: Large Language Models are not knowledge bases. Instead, they are probabilistic models of knowledge bases

Analogy: Databases versus Statistical Models of Databases

Large Language Models : Knowledge Bases :: Statistical DB Models : Databases

Statistical models of databases:

- Data cleaning
  - A person with age “2023” is probably an error
- Query Optimization
  - Estimate the sizes of intermediate tables when executing a query plan

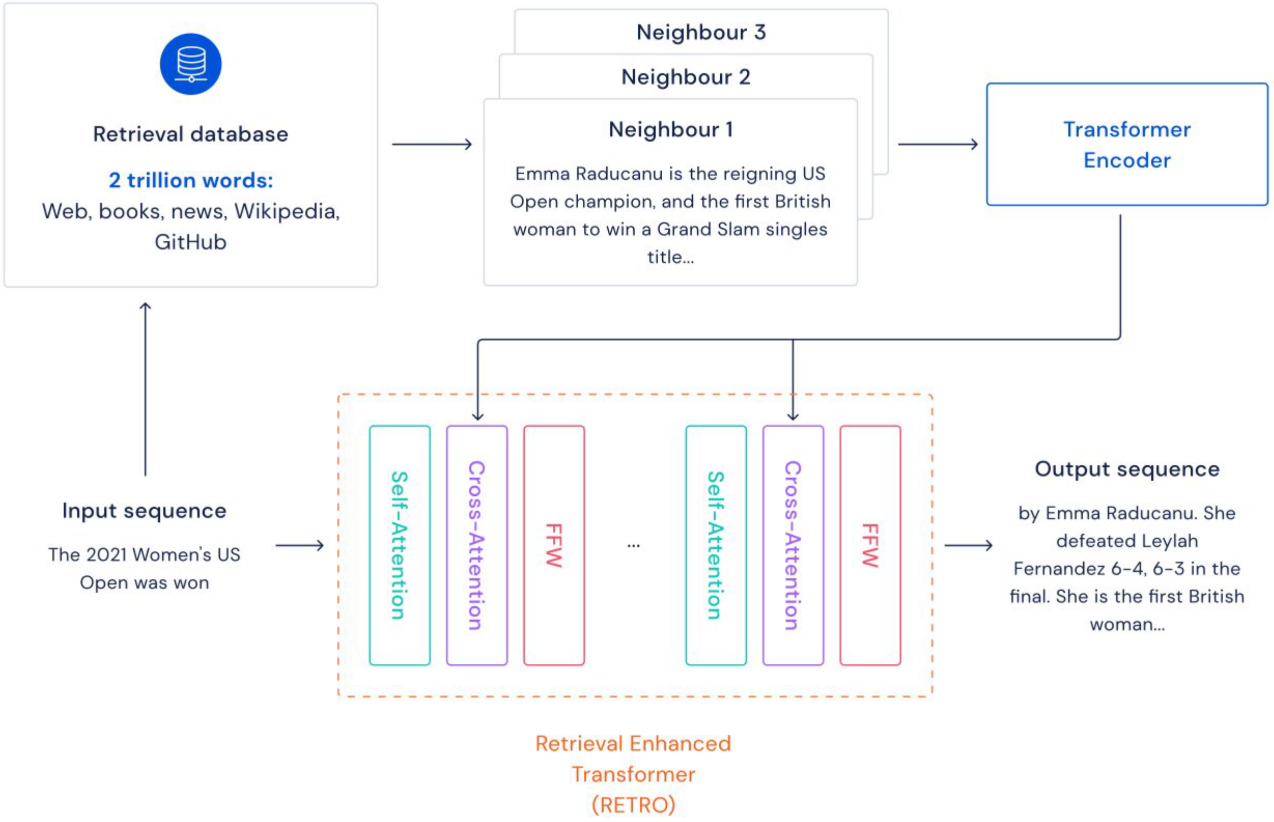
ID	Name	State
49283	Phil Knight	Oregon
33924	Mark Zuckerberg	California
42238	Sundar Pichai	California
88499	Marc Benioff	California

Query: What state does Karen Lynch work in?
Database system:
Unknown
Probabilistic model:
California (75%)
Oregon (25%)
Correct answer:
Rhode Island

We want knowledge bases, not statistical models of knowledge bases

Current Efforts to Address Problems: Retrieval-Augmented LMs

- Retrieval-Augmented Language Models
  - Use input sequence to search external document collections or knowledge graphs
  - Fuse results with the query to generate the answer
  - Bing probably implements this
- Benefits
  - Network can be 10x smaller (RETRO)
  - External documents can be updated without retraining
  - Reduces hallucination
  - Answer can be attributed to source documents
- Issues
  - Implicit world knowledge (in LLM) can interfere with knowledge from retrieved documents to cause hallucinations
  - Evaluations (Bing, NeevaAI, perplexity.ai, YouChat) show 48.5% of generated sentences are not fully supported by retrieved documents and 25.5% of cited documents are irrelevant (Liu, et al. 2023)
  - Vulnerable to poisoning of external knowledge sources

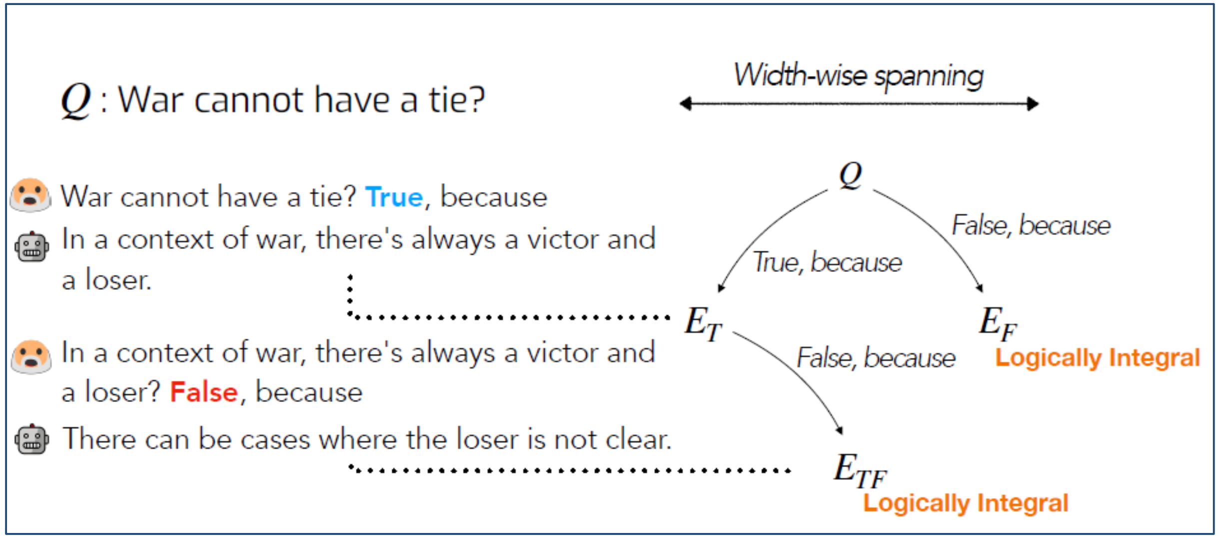


RETRO: Borgeaud, et al. 2021; 2022

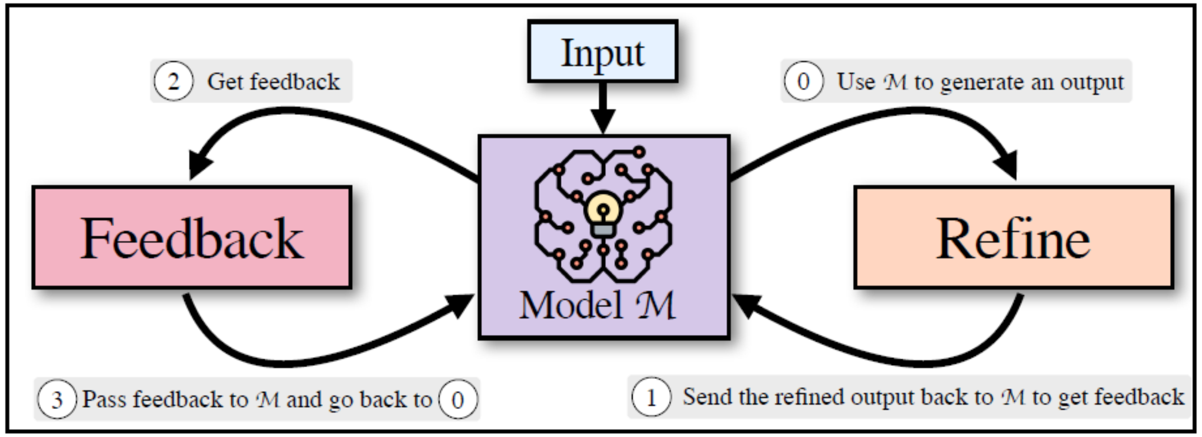
Improving Consistency

- Ask multiple, logically-related questions and apply MaxSAT solver to find the most coherent belief

- Self-Refinement: Ask model to critique and refine its own output



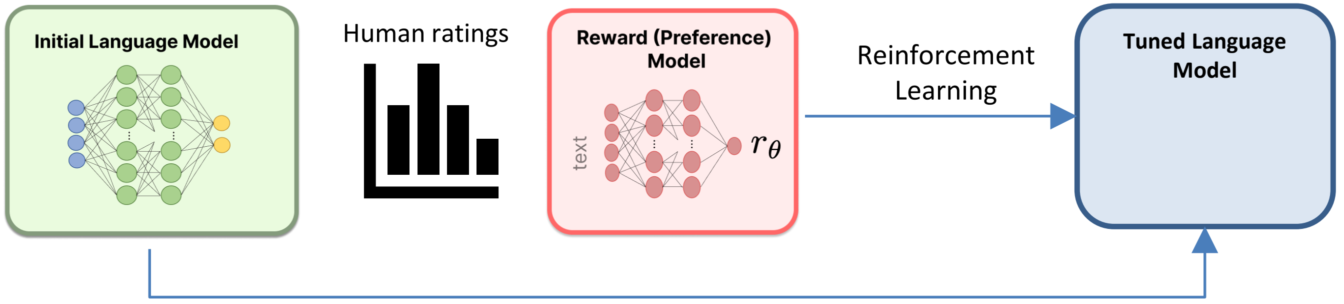
Bhagavatula, et al, 2022



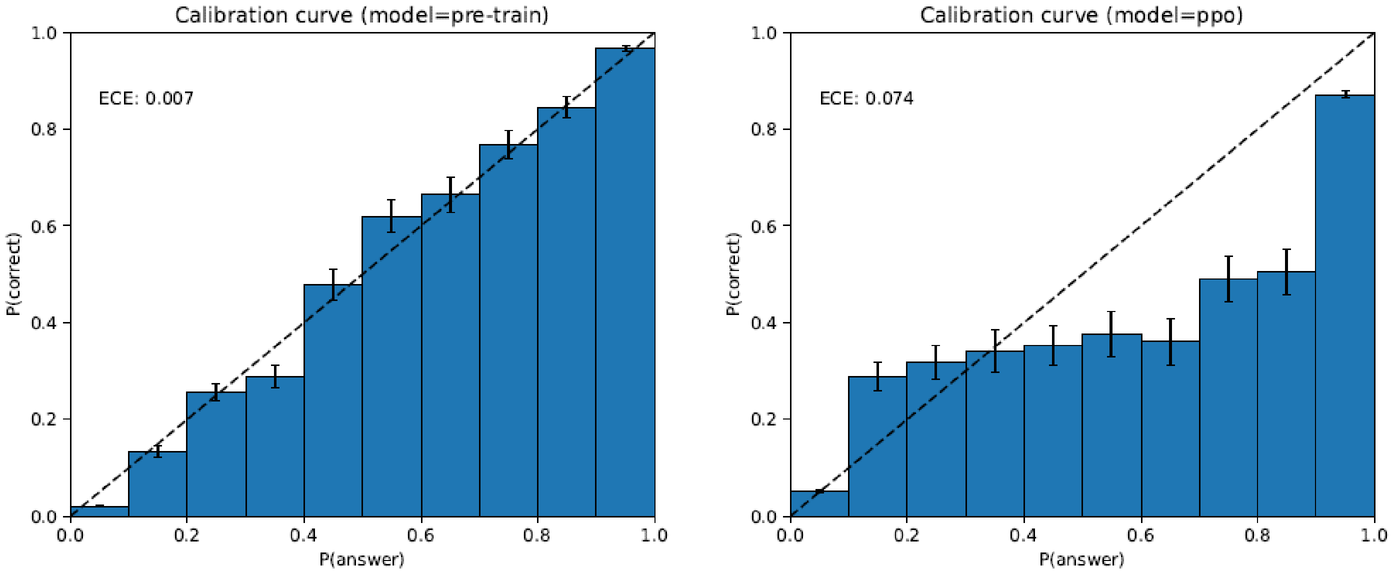
Madaan, et al., 2023

Reducing Dangerous and Socially Inappropriate Outputs

- Reinforcement-learning from human feedback
  - Step 1: Collect feedback on suitability of generated output
  - Step 2: Train a reward model (preference model)
  - Step 3: Tune the language model via reinforcement learning to maximize the reward while changing probabilities as little as possible
- Shortcomings
  - Reduces, but does not eliminate toxic and dangerous outputs
  - Definition of “inappropriate” will reflect human biases and is not inspectable; leads to political controversy
  - RLHF seriously damages output calibration
- Future Steps
  - Train a second language model to recognize inappropriate content
  - Constitutional AI (Bai, et al. 2023)



Modified from HuggingFace



GPT-4 Calibration Curves

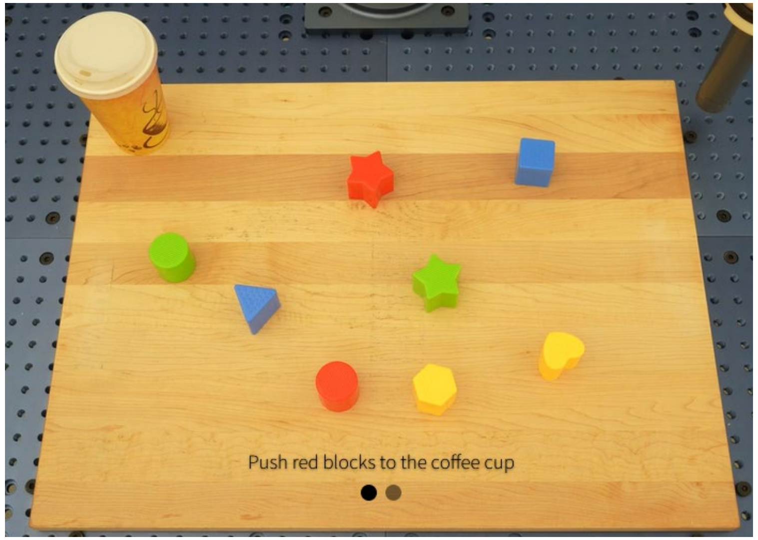
Learning and Applying Non-Linguistic Knowledge

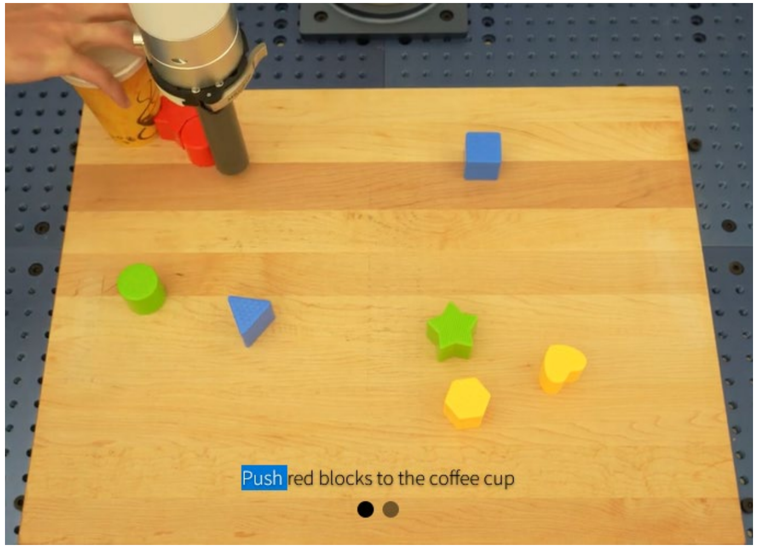
Multi-modal networks

- Kosmos-1, Flamingo: Trained on text and images. Strong few-shot learning capability on image tasks
- PaLM-E: Trained on text, images, state estimation, and robot actions. Output: text, robot commands.
- Main focus: Few-shot learning for vision-language tasks

Calling out to external tools

- ToolFormer: Learn to invoke APIs for calendar, web search, calculator
- ChatGPT Plugins
- Adept.com: "automate any software process" (email, Salesforce, Google sheets, shopping)

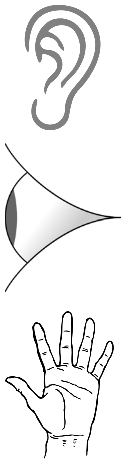




WHAT WE SHOULD BE DOING INSTEAD

Modular AI Systems

Neuroscience suggests that separate brain regions are responsible for each of these functions



Planning

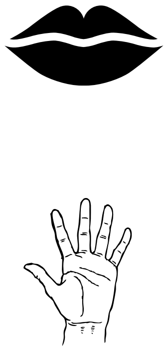
Meta-Cognition
Self-Monitoring
Orchestration

Formal Reasoning

Language understanding & generation

Common sense knowledge

Factual world knowledge



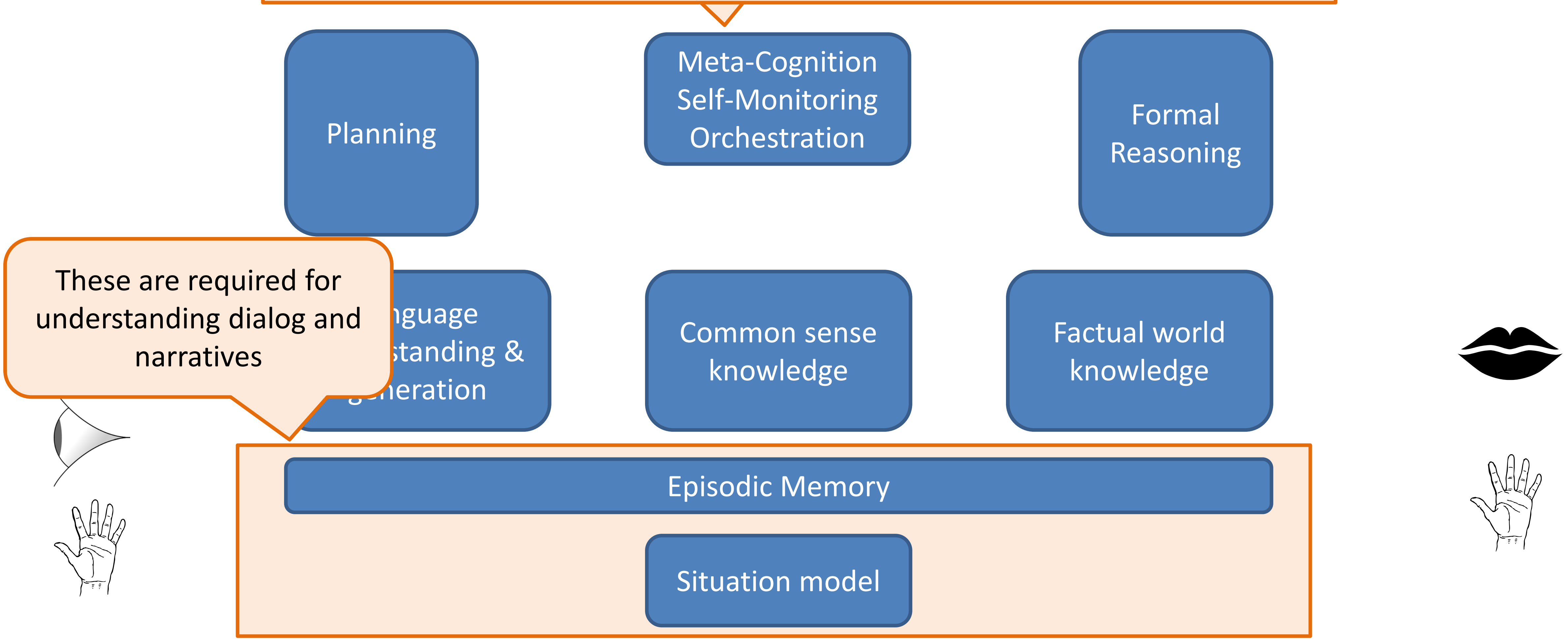
Episodic Memory

Situation model

Mahowald, et al. 2023 “Dissociating language and thought in large language models: a cognitive perspective.”

Beyond Lar

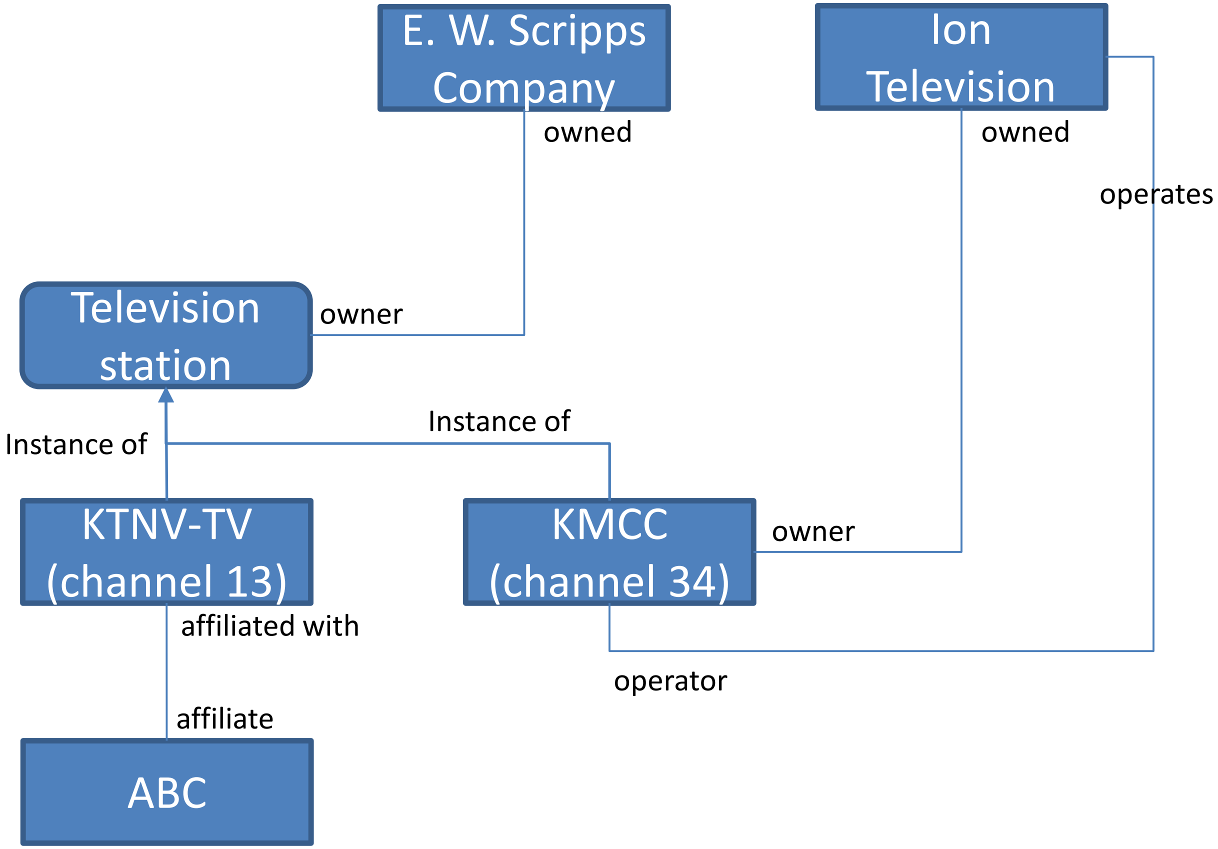
Adding missing modules and disentangling factual world knowledge from language and common sense could address virtually all of the shortcomings of today's LLMs



Representing Factual World Knowledge as a Knowledge Graph

https://en.wikipedia.org/wiki/KTNV-TV:

"**KTNV-TV** (channel 13) is a television station in Las Vegas, Nevada, United States, affiliated with ABC. It is owned by the E. W. Scripps Company alongside Laughlin-licensed Ion Television owned-and-operated station KMCC (channel 34)."



End-to-End Training for Factual Knowledge and Dialogue

- Separate Language Skill from Factual World Knowledge
- Represent world knowledge as a knowledge graph over an extensible ontology

- Architecture:
  - Encoder:
    - Given:
      - Paragraph
    - Find:
      - Set of relevant facts in the knowledge graph (adding them if necessary)
      - Communicative goal and other pragmatic information
  - Decoder:
    - Given:
      - Set of relevant facts in the knowledge graph
      - Communicative goal and other pragmatic information
    - Output:
      - Paragraph
  - Train to reproduce the input paragraph

Previous effort: NELL

- Never-Ending Learning (Mitchell, et al. 2015)
  - Extracted triples
  - Collected and integrated evidence in favor of and against each triple
  - Extended its initial ontology
  - Inferred new relationships and their arguments (and argument restrictions)
- Ran from 2010-2018
- Is it time for another NELL, but using LLMs?

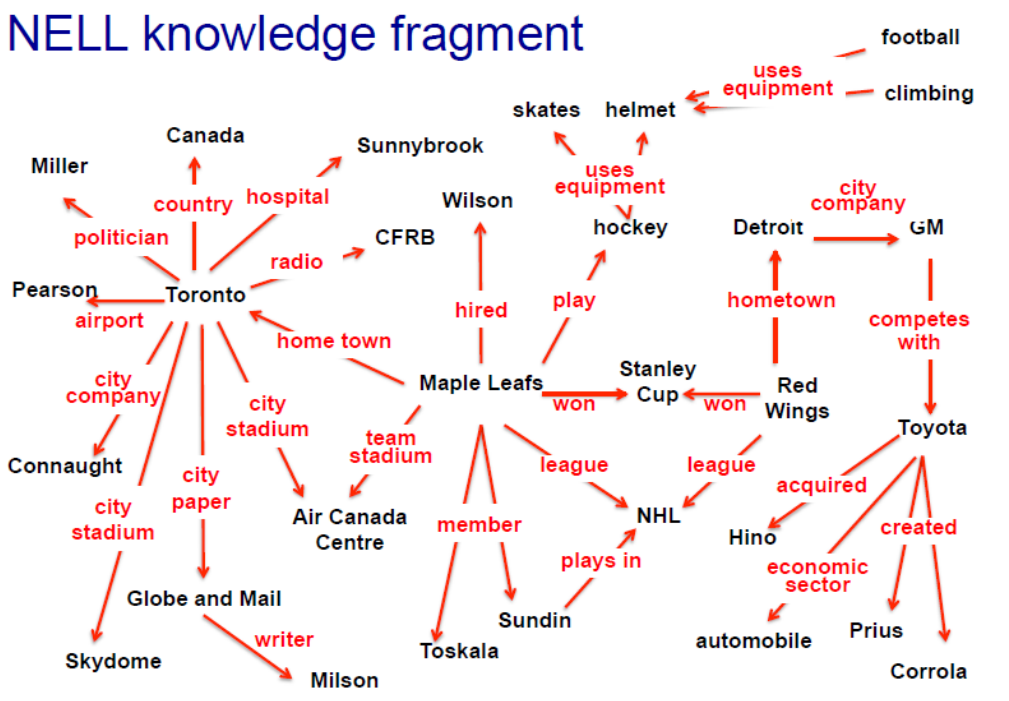


Figure 1: Fragment of the 80 million beliefs NELL has read from the web. Each edge represents a belief triple (e.g., play(MapleLeafs, hockey), with an associated confidence and provenance not shown here. This figure contains only correct beliefs from NELL's KB – it has many incorrect beliefs as well since NELL is still learning.

Can we bootstrap from ChatGPT? (May 12 version)

Prompt: Read the following paragraph and list all of the simple facts that it contains: "KTNV-TV (channel 13) is a television station in Las Vegas, Nevada, United States, affiliated with ABC. It is owned by the E. W. Scripps Company alongside Laughlin-licensed Ion Television owned-and-operated station KMCC (channel 34)." Here are two example facts: "KTNV-TV is a television station" and "KTNV-TV is located in Las Vegas, Nevada, United States".

- KTNV-TV is a television station.
- KTNV-TV is located in Las Vegas, Nevada, United States.
- KTNV-TV is affiliated with ABC.
- KTNV-TV is owned by the E. W. Scripps Company.
- KMCC (channel 34) is an Ion Television owned-and-operated station.
- KMCC is licensed in Laughlin.

May need additional prompt engineering to separate the facts

May need to verify that each fact was present in the paragraph

Apply ideas from NELL to fuse these facts into a well-supported knowledge graph

Related Work

- Extracting knowledge graphs from LLMs
  - Develop various prompting and fill-in-the-blank tasks to extract KG tuples
    - Petroni, et al. 2019 “Language models as knowledge bases?”
- Applying LLMs to construct knowledge graphs from documents
  - Given nodes from an existing KG, extract relations by processing the corpus using an LLM
    - Wang, et al. 2020 “Language models are open knowledge graphs”
  - Extract nodes from a set of documents using an LLM. Then apply a classifier to predict whether an edge exists
    - Melnyk, et al. 2022 “Knowledge Graph Generation from Text”

End-to-End Training for Next Phrase Prediction

- Encoder:
  - Given:
    - conversation so far
  - Build the situation model:
    - goals of the speaker
    - beliefs and arguments of the speaker
    - how the conversation implements a narrative plan
    - facts asserted thus far
- Decoder:
  - Given:
    - goals and beliefs of the speaker
    - conversation and situation model thus far
  - Do:
    - extend the narrative plan
    - retrieve relevant knowledge from the knowledge graph
    - generate the next phrase

Attaining Truthfulness

- The knowledge graph approach assumes there is a single, coherent, true model of the world
  - People disagree on the truth
  - Existing scientific evidence may not be conclusive
  - There are cultural variations
- Possible approaches
  - Build internally-coherent micro-worlds
  - Support each assertion with an argument from evidence
- Our AI systems need to be able to reason about the trustworthiness of information sources
  - Google has a whole team dedicated to rating the trustworthiness of web sites
  - This has been a continual battle between spammers and the search engines
  - It will get worse with the advent of LLM-based systems

Missing Aspects and Open Questions

- Missing forms of knowledge
  - General rules that are difficult to capture as knowledge graph triples
  - Actions that can be taken in the world
    - preconditions
    - results and side-effects
    - costs
  - Ongoing processes
    - water flowing or filling a container
    - battery discharging
- Meta-cognitive subsystem
  - Self-monitoring for social acceptability
  - Self-monitoring for ethical appropriateness
  - Orchestration of planning, reasoning, memory, and language

Summary

- Existing LLMs have many flaws
  - They are statistical models of knowledge bases rather than knowledge bases
  - They are expensive to update with new/changing factual knowledge
  - They produce socially and ethically unacceptable outputs
- We should be building modular AI systems that
  - separate linguistic skill from world knowledge
  - marshall planning, reasoning, and knowledge to build situation models of narratives/dialogues
  - record and retrieve from episodic memory
  - create and update world knowledge
- There are many, many details to be worked out!!

References

- Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., Mckinnon, C., Dec, C. L., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-johnson, E., … Kaplan, J. (2022). Constitutional AI: Harmlessness from AI Feedback. ArXiv, 2212.08073(v1).
- Bhagavatula, C., Hwang, J. D., Downey, D., Bras, R. Le, Lu, X., Sakaguchi, K., Swayamdipta, S., West, P., & Choi, Y. (2022). I2D2: Inductive Knowledge Distillation with NeuroLogic and Self-Imitation. *ArXiv*, *2212.0924*(v1). https://arXiv.org/abs/2212.0924
- Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Driessche, G. van den, Lespiau, J.-B., Damoc, B., Clark, A., Casas, D. de Las, Guy, A., Menick, J., Ring, R., Hennigan, T., Huang, S., Maggiore, L., Jones, C., Cassirer, A., … Sifre, L. (2021). Improving language models by retrieving from trillions of tokens. *ArXiv*, *2112.04426*(v3). http://arxiv.org/abs/2112.04426
- Goldberg, Y. (2018). Assessing BERT’s Syntactic Abilities. *ArXiv*, *1901.05287*(v1), 2–5. https://arXiv.org/abs/1901.05287
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, *55*(12), 1–38. https://doi.org/10.1145/3571730
- Jung, J., Qin, L., Welleck, S., Brahman, F., Bhagavatula, C., Bras, R. Le, & Choi, Y. (2020). Maieutic Prompting: Logically Consistent Reasoning with Recursive Explanations. *ArXiv*, *2205.11822*. https://arXiv.org/abs/2205.11822
- Khattab, O., Santhanam, K., Li, X. L., Hall, D., Liang, P., Potts, C., & Zaharia, M. (2022). Demonstrate-Search-Predict: Composing retrieval and language models for knowledge-intensive NLP. *ArXiv*, *2212.14024*(v2). https://arXiv.org/abs/2212.14024
- Knight, W. (2023). OpenAI’s CEO Says the Age of Giant AI Models Is Already Over. *Wired*, 17 April 2023.

References (2)

- Liu, N. F., Zhang, T., & Liang, P. (2023). Evaluating Verifiability in Generative Search Engines. *ArXiv, 2304.09848*(v1). https://arxiv.org/abs/2304.09848
- Liu, Q., Yogatama, D., Blunsom, P. (2022) Relational Memory Augmented Language Models. *ArXiv, 2201.09680*. https://arxiv.org/abs/2201.09680
- Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). Dissociating language and thought in large language models: a cognitive perspective. *ArXiv*, *2301.06627*(v1). https://arxiv.org/abs/2301.06627
- Marvin, R., & Linzen, T. (2018). Targeted Syntactic Evaluation of Language Models. *EMNLP*, 1192–1202. https://aclanthology.org/D18-1151/
- Melnyk, I., Dognin, P., & Das, P. (2022). Knowledge Graph Generation From Text. *Findings of the Association for Computational Linguistics: EMNLP 2022*, 1610–1622.
- Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A., Grave, E., LeCun, Y., & Scialom, T. (2023). Augmented Language Models: a Survey. *ArXiv*, *2302.07842*(v1), 1–33. http://arxiv.org/abs/2302.07842
- Mitchell, T., Cohen, W., Hruschka, E., Talukdar, P., Yang, B., Betteridge, J., Carlson, A., Dalvi, B., Gardner, M., Kisiel, B., Krishnamurthy, J., Lao, N., Mazaitis, K., Mohamed, T., Nakashole, N., Platanios, E., Ritter, A., Samadi, M., Settles, B., … Welling, J. (2018). Never-ending learning. *Communications of the ACM*, *61*(5), 103–115. https://doi.org/10.1145/3191513

References (3)

- Mok, A. (2023). ChatGPT could cost over $700,000 per day to operate. Microsoft is reportedly trying to make it cheaper. *Insider*, April 20, 2023.
- Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., & Riedel, S. (2019). Language models as knowledge bases? *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2463–2473. https://doi.org/10.18653/v1/d19-1250
- Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., & Shoham, Y. (2023). In-Context Retrieval-Augmented Language Models. *ArXiv*, *2302.00083*(v1). http://arxiv.org/abs/2302.00083
- Ramezanil, A., Valasquez, G. A., Rasuli, S., & Knowles, L. R. (2018). Neuroanatomical and Neurocognitive Functions of the Structure of the Mind: Clinical and Teaching Implications. *Current Opinions in Neurological Science*, *2*(6). https://scientiaricerca.com/srcons/SRCONS-02-00079.php
- Russin, J., O'Reilly, R. C., & Bengio, Y. (2020). Deep Learning Needs a Prefrontal Cortex. *Bridging AI and Cognitive Science (ICLR 2020 Workshop)*, 1–11. https://arxiv.org/abs/1910.00744
- Schick, T., Lomeli, M., Dwivedi-yu, J., & Dessì, R. (2022). Toolformer: Language Models Can Teach Themselves to Use Tools. *ArXiv*, *2302.04761*(v1). https://arXiv.org/abs/2302.04761
- Wang, C., Liu, X., & Song, D. (2020). Language Models are Open Knowledge Graphs. *ArXiv*, *2010.11967*(v1), 1–30. http://arxiv.org/abs/2010.11967