What is a Language Model and Why Is It Not Capable of Having Intentions?

Andrey Kutuzov
Language Technology Group
University of Oslo

AIAI
16 May 2024

Where I come from?

- LTG: Language Technology Group
- Section for Machine Learning,
  Department of Informatics, University of Oslo
- Run our own study programs (BSc + MSc)
- $\sim$4 permanent, 2 adjuncts, 3 postdocs,
  2 researchers, 8 PhDs
- Natural Language Processing (NLP):
- also known as 'computational linguistics'

Where I come from?

- LTG: Language Technology Group
- Section for Machine Learning, Department of Informatics, University of Oslo
- Run our own study programs (BSc + MSc)
- ∼4 permanent, 2 adjuncts, 3 postdocs, 2 researchers, 8 PhDs
- Natural Language Processing (NLP):
- also known as 'computational linguistics'
- ...and of course we train and evaluate large language models (for English and Norwegian)

`https://www.mn.uio.no/ifi/english/research/groups/ltg/`

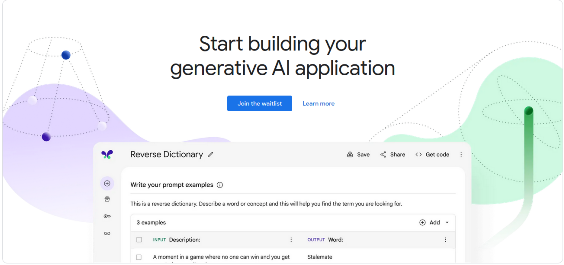
Contents

What are language models?

(ChatGPT, a generative language model by OpenAI)
`https://chatgpt.com/`

What are language models?

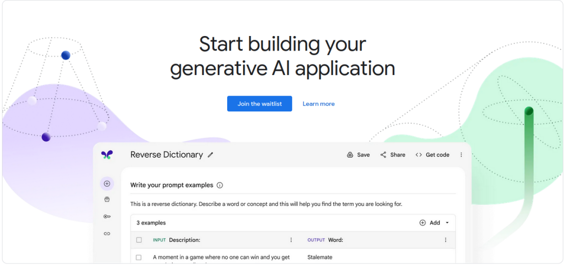
(ChatGPT, a generative language model by OpenAI)
`https://chatgpt.com/`



(PaLM 2, a generative language model announced by Google in May 2023)
`https://ai.google/discover/palm2/`

What are language models?

(*ChatGPT, a generative language model by OpenAI*)
`https://chatgpt.com/`

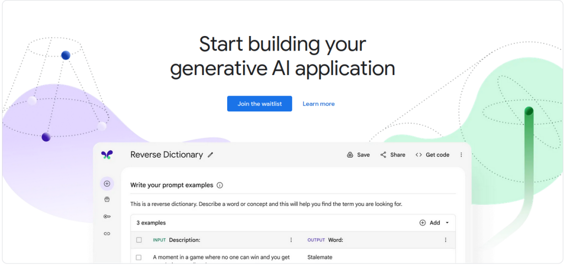


(*PaLM 2, a generative language model announced by Google in May 2023*)
`https://ai.google/discover/palm2/`

Are these ‘language models’ artificial intelligence (AI)?

What are language models?

(*ChatGPT, a generative language model by OpenAI*)
`https://chatgpt.com/`



(*PaLM 2, a generative language model announced by Google in May 2023*)
`https://ai.google/discover/palm2/`

Are these 'language models' artificial intelligence (AI)? And what do they actually 'model'?

What are language models?

Roughly speaking, language modeling is

...predicting the next word in the text given the previous words

What are language models?

Roughly speaking, language modeling is

...predicting the next word in the text given the previous words

For example

- ‘What is the meaning of <PREDICT>’....

What are language models?

Roughly speaking, language modeling is

...predicting the next word in the text given the previous words

For example

- ‘What is the meaning of <PREDICT>’....
- hmm... ‘life’?

What are language models?

Roughly speaking, language modeling is

...predicting the next word in the text given the previous words

For example

- 'What is the meaning of <PREDICT>'....
- hmm... 'life'?
- Yes! 'What is the meaning of life'.

What are language models?

Roughly speaking, language modeling is

...predicting the next word in the text given the previous words

For example

- 'What is the meaning of <PREDICT>'....
- hmm... 'life'?
- Yes! 'What is the meaning of life'.
- 'She is a researcher in natural language <PREDICT>'....

What are language models?

Roughly speaking, language modeling is

...predicting the next word in the text given the previous words

For example

- ‘What is the meaning of <PREDICT>’....
- hmm... ‘life’?
- Yes! ‘What is the meaning of life’.
- ‘She is a researcher in natural language <PREDICT>’....
- hmm... ‘processing’?

What are language models?

Roughly speaking, language modeling is

...predicting the next word in the text given the previous words

For example

- 'What is the meaning of <PREDICT>'....
- hmm... 'life'?
- Yes! 'What is the meaning of life'.
- 'She is a researcher in natural language <PREDICT>'....
- hmm... 'processing'?
- No! 'She is a researcher in natural language understanding'.

What are language models?

Roughly speaking, language modeling is

...predicting the next word in the text given the previous words

For example

- ‘What is the meaning of <PREDICT>’....
- hmm... ‘life’?
- Yes! ‘What is the meaning of life’.
- ‘She is a researcher in natural language <PREDICT>’....
- hmm... ‘processing’?
- No! ‘She is a researcher in natural language understanding’.

- Idea dates back to [Shannon, 1948]
- actively used since the 1980s for Machine Translation and Automated Speech Recognition
- $\sim$10 years ago, with neural LMs, became central in NLP and more.

What are language models?

Language modelling as two tasks

- Task 1: to estimate probabilities of natural language sequences:

What are language models?

Language modelling as two tasks

- Task 1: to estimate probabilities of natural language sequences:
  - 'What is the probability of *lazy dog*?'

What are language models?

Language modelling as two tasks

- Task 1: to estimate probabilities of natural language sequences:
  - 'What is the probability of *lazy dog*?'
  - 'What is the probability of *The quick brown fox jumps over the lazy dog*?'

What are language models?

Language modelling as two tasks

- Task 1: to estimate probabilities of natural language sequences:
  - 'What is the probability of *lazy dog*?'
  - 'What is the probability of *The quick brown fox jumps over the lazy dog*?'
  - 'What is the probability of *green colorless ideas sleep furiously*?'

What are language models?

Language modelling as two tasks

- Task 1: to estimate probabilities of natural language sequences:
  - 'What is the probability of *lazy dog*?'
  - 'What is the probability of *The quick brown fox jumps over the lazy dog*?'
  - 'What is the probability of *green colorless ideas sleep furiously*?'
- Task 2: to estimate the probability of a word x to follow a word sequence S of length n:

What are language models?

Language modelling as two tasks

- Task 1: to estimate probabilities of natural language sequences:
  - 'What is the probability of *lazy dog*?'
  - 'What is the probability of *The quick brown fox jumps over the lazy dog*?'
  - 'What is the probability of *green colorless ideas sleep furiously*?'
- Task 2: to estimate the probability of a word x to follow a word sequence S of length n:
  - 'What is the probability of seeing *jumps* after *The quick brown fox*?'

What are language models?

Language modelling as two tasks

- Task 1: to estimate probabilities of natural language sequences:
  - 'What is the probability of *lazy dog*?'
  - 'What is the probability of *The quick brown fox jumps over the lazy dog*?'
  - 'What is the probability of *green colorless ideas sleep furiously*?'
- Task 2: to estimate the probability of a word x to follow a word sequence S of length n:
  - 'What is the probability of seeing *jumps* after *The quick brown fox*?'
- These two are closely related, almost the same task:

$$P(w_{1:n}) = P(w_1)P(w_2|w_1)P(w_3|w_{1:2})P(w_4|w_{1:3})...P(w_n|w_{1:n-1}) \tag{1}$$

- Any system able to yield $P(x)$ given S is a language model (LM).

What are language models?

Language modelling as two tasks

- Task 1: to estimate probabilities of natural language sequences:
  - 'What is the probability of *lazy dog*?'
  - 'What is the probability of *The quick brown fox jumps over the lazy dog*?'
  - 'What is the probability of *green colorless ideas sleep furiously*?'
- Task 2: to estimate the probability of a word x to follow a word sequence S of length n:
  - 'What is the probability of seeing *jumps* after *The quick brown fox*?'
- These two are closely related, almost the same task:

$$P(w_{1:n}) = P(w_1)P(w_2|w_1)P(w_3|w_{1:2})P(w_4|w_{1:3})...P(w_n|w_{1:n-1}) \quad (1)$$

- Any system able to yield $P(x)$ given S is a language model (LM).

Computational language models are data-driven: they are trained to learn the probabilities from large natural text collections.

Evaluation of language models

'She is a researcher in natural language...

Evaluation of language models

'She is a researcher in natural language... snow-boarding'?!
I am perplexed!

Evaluation of language models

'She is a researcher in natural language... snow-boarding'?!
I am perplexed!

- One can evaluate and compare LMs by their perplexity:
  - how perplexed/surprised is the model by test word sequences
  - the lower the better.

Evaluation of language models

'She is a researcher in natural language... snow-boarding'?!
I am perplexed!

- One can evaluate and compare LMs by their perplexity:
  - how perplexed/surprised is the model by test word sequences
  - the lower the better.
- For each of i words in the test corpus, find how probable it is according to the LM:

$$ENTROPY_i = -\log_2 LM\left(w_i \middle| w_{1:i-1}\right)$$

Evaluation of language models

'She is a researcher in natural language... snow-boarding'?!

I am perplexed!

- One can evaluate and compare LMs by their perplexity:
  - how perplexed/surprised is the model by test word sequences
  - the lower the better.
- For each of i words in the test corpus, find how probable it is according to the LM:

$$\begin{aligned} ENTROPY_i &= -\log_2 LM\left(w_i \middle| w_{1:i-1}\right) \\ PERPLEXITY_i &= 2^{ENTROPY_i} \end{aligned} \tag{2}$$

Evaluation of language models

'She is a researcher in natural language... snow-boarding'?!
I am perplexed!

- One can evaluate and compare LMs by their perplexity:
  - how perplexed/surprised is the model by test word sequences
  - the lower the better.
- For each of i words in the test corpus, find how probable it is according to the LM:

$$\begin{aligned} ENTROPY_i &= -\log_2 LM\left(w_i \middle| w_{1:i-1}\right) \\ PERPLEXITY_i &= 2^{ENTROPY_i} \end{aligned} \tag{2}$$

- exponentiated negative log-likelihoods per token
- For corpus perplexity, you simply average token perplexities.

What are language models?

Any language model is a text generator by definition

What are language models?

Any language model is a text generator by definition

Autoregressive or causal generation:

- feed a word or a sentence (prompt) into the LM
- get a probability distribution over what words are likely to come next
- pick the most probable word from this distribution (or use some form of sampling)
- feed it right back in the LM together with the previous words
- repeat this process and you're generating text!

Slightly rephrasing `https://karpathy.github.io/2015/05/21/rnn-effectiveness/`

What are language models?

Any language model is a text generator by definition

Autoregressive or causal generation:

- feed a word or a sentence (prompt) into the LM
- get a probability distribution over what words are likely to come next
- pick the most probable word from this distribution (or use some form of sampling)
- feed it right back in the LM together with the previous words
- repeat this process and you're generating text!

Slightly rephrasing `https://karpathy.github.io/2015/05/21/rnn-effectiveness/`

This is what ChatGPT or GPT-4 do. Thus, generative language models.
Generating word sequences to pretend as best as they can that these sequences are generated by humans: 'Imitation Game'.

What are language models?

Any language model is a text generator by definition

Autoregressive or causal generation:

- feed a word or a sentence (prompt) into the LM
- get a probability distribution over what words are likely to come next
- pick the most probable word from this distribution (or use some form of sampling)
- feed it right back in the LM together with the previous words
- repeat this process and you're generating text!

Slightly rephrasing `https://karpathy.github.io/2015/05/21/rnn-effectiveness/`

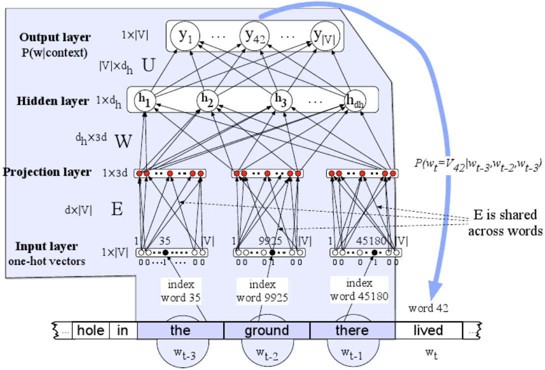
This is what ChatGPT or GPT-4 do. Thus, generative language models.
Generating word sequences to pretend as best as they can that these sequences are generated by humans: 'Imitation Game'.
Decide for yourself whether this counts as 'AI'.

Contents

What created modern ‘Generative AI’ hype?

Modern language models are built with multi-layered artificial neural networks

- First neural LM in [Bengio et al., 2003] used feed-forward neural network architecture



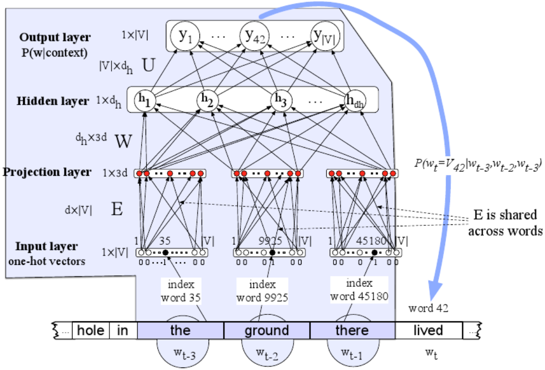
- produced word representations (embeddings) as a by-product in its hidden layers.

(image from Jurafsky and Martin, 2023)

What created modern ‘Generative AI’ hype?

Modern language models are built with multi-layered artificial neural networks

- First neural LM in [Bengio et al., 2003] used feed-forward neural network architecture



- produced word representations (embeddings) as a by-product in its hidden layers.

(image from Jurafsky and Martin, 2023)

But things have moved forward since then. In what ways?

1. Increased compute

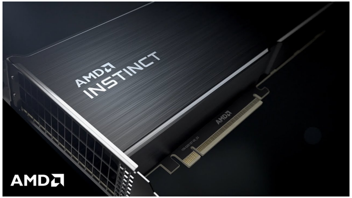
- Hardware capabilities are growing: graphic processing units (GPUs) and Tensor Processing Units (TPUs). They excel in parallelized matrix multiplication.

1. Increased compute

- Hardware capabilities are growing: graphic processing units (GPUs) and Tensor Processing Units (TPUs). They excel in parallelized matrix multiplication.
- Compute divide: who can afford burning 100K GPU/hours to train a GPT-10B model for a mid-sized language?

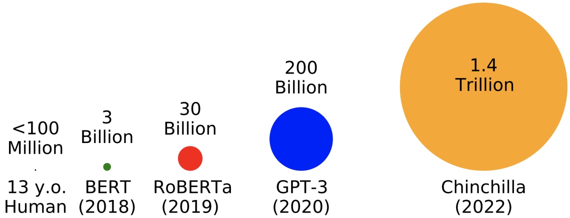
1. Increased compute

- Hardware capabilities are growing: graphic processing units (GPUs) and Tensor Processing Units (TPUs). They excel in parallelized matrix multiplication.
- Compute divide: who can afford burning 100K GPU/hours to train a GPT-10B model for a mid-sized language?



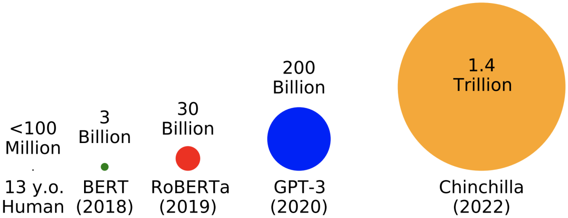
2. Increased data

LMs are trained on raw texts: lots of data to crawl from the Internet (most of it in English).
Training corpora sizes for some famous LMs in running words:



2. Increased data

LMs are trained on raw texts: lots of data to crawl from the Internet (most of it in English). Training corpora sizes for some famous LMs in running words:



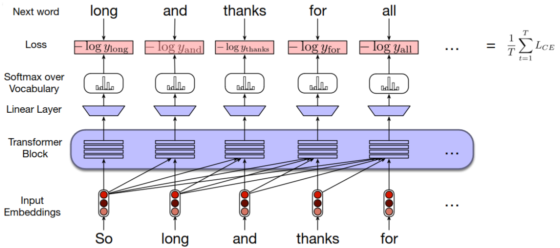
- Formal linguistic skills of language models improve a lot when the size of the training data increases
- ...unlike functional communicative competence (social reasoning, pragmatics, etc), which often require special modules.

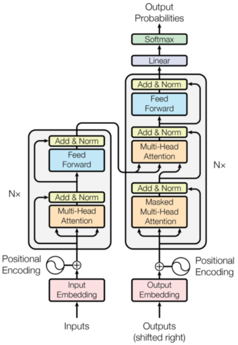
3. Better architectures: transformers

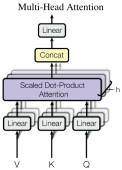
Transformer

- A sequence of feedforward layers
- multi-headed self-attention
- positional encoding

Transformers allowed to use the existing data and compute in the most optimal way.







(image from Jurafsky and Martin, 2023)

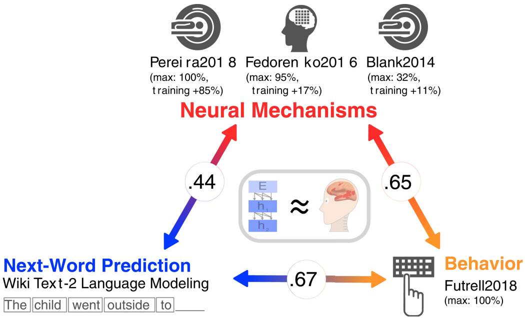
Contents

Language models similar to human brain?

Predictive language processing in humans

- ‘Models that perform better at predicting the next word in a sequence also better predict brain measurements’
- ‘predictive processing fundamentally shapes the language comprehension mechanisms in the brain’

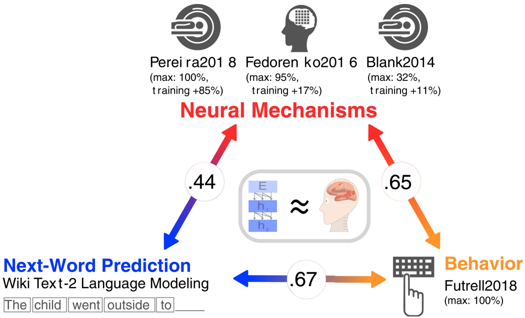
[Schrimpf et al., 2021]



Language models similar to human brain?

Predictive language processing in humans

- ‘Models that perform better at predicting the next word in a sequence also better predict brain measurements’
- ‘predictive processing fundamentally shapes the language comprehension mechanisms in the brain’

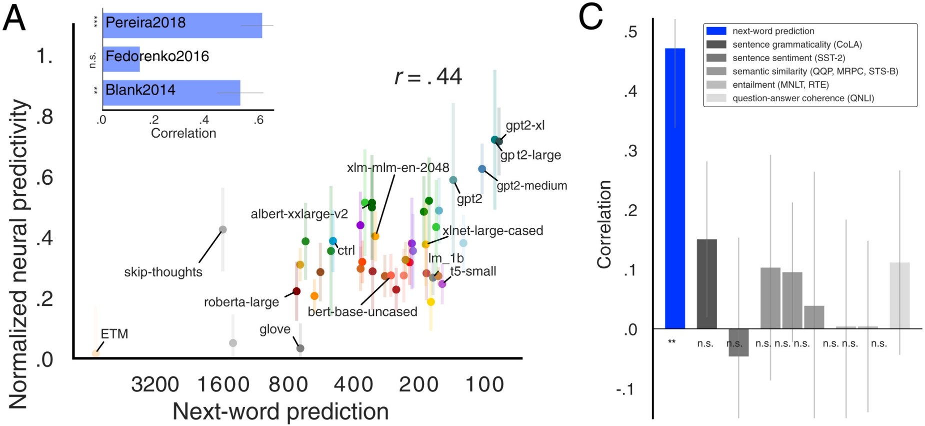


[Schrimpf et al., 2021]

Human language system and computational LMs are both optimized to predict upcoming words for efficient meaning extraction?
Might be!

Language models similar to human brain?

Interestingly, it is specifically next word prediction performance that correlates with human language processing activities (not other NLP tasks):



[Schrimpf et al., 2021]

Language models similar to human brain?

Tool to study human language processing?

- Neural LMs are much better correlated with brain data than the previous-generation LMs.
- They are not exactly models of brain, but their architectures capture important properties of language processing in humans.

'It seems that language modeling encourages a neural network to build a joint probability model of the linguistic signal, which implicitly requires sensitivity to diverse kinds of regularities in the signal'

[Schrimpf et al., 2021]

Language models similar to human brain?

Tool to study human language processing?

- Neural LMs are much better correlated with brain data than the previous-generation LMs.
- They are not exactly models of brain, but their architectures capture important properties of language processing in humans.

'It seems that language modeling encourages a neural network to build a joint probability model of the linguistic signal, which implicitly requires sensitivity to diverse kinds of regularities in the signal'

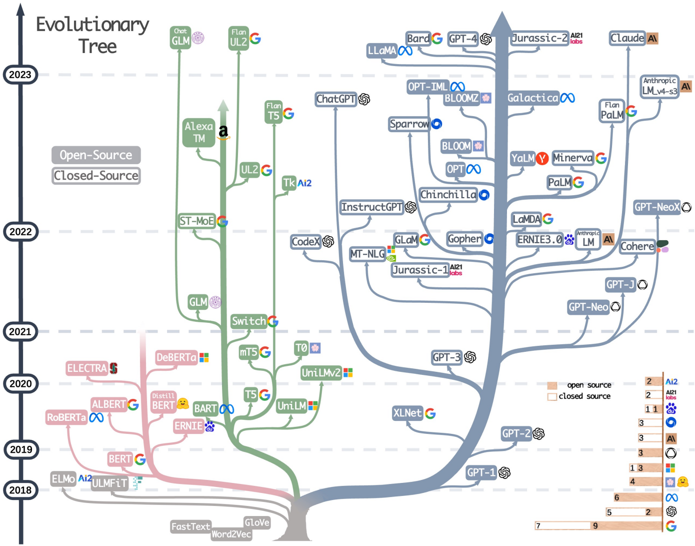
[Schrimpf et al., 2021]

NB: LLMs are much worse with functional tasks (e.g. related to theory of mind)!

[Mahowald et al., 2024]

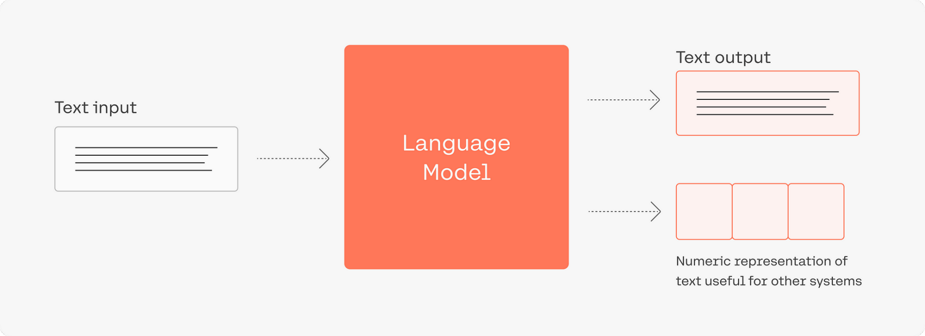
Contents

Modern large language models



https://github.com/Mooler0410/LLMsPracticalGuide

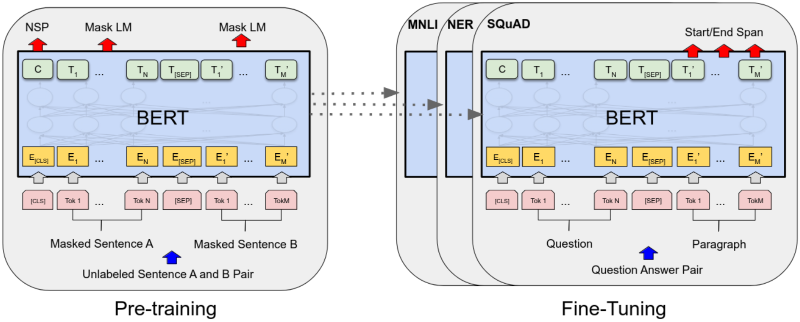
Modern large language models



There are three major types of modern LMs aimed at producing different outputs: encoder-only, decoder-only and encoder-decoder.

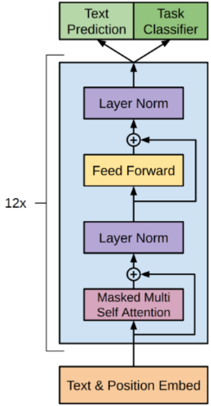
1. Encoder language models

- Trained to produce useful representations of input words / sequences (encode them)
- also known as masked language models
- popular example: BERT [Devlin et al., 2019]
- not used much for generation, but excel in classification, etc



2. Decoder language models

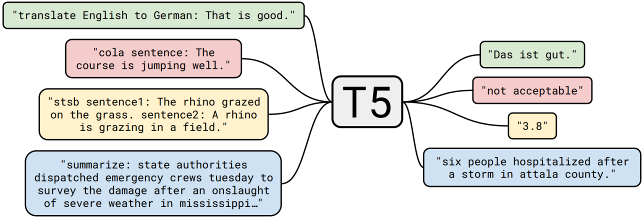
- Trained to predict the next word based on the previous words
- decoding the current model state into human language words
- also known as autoregressive or causal models
- excel in text generation
- most classical type of language models, dating back 70 years
- popular examples: GPT-3 [Brown et al., 2020], ChatGPT, GPT-4, Mistral [Jiang et al., 2023] and what not.



Architectures

3. Encoder-decoder language models

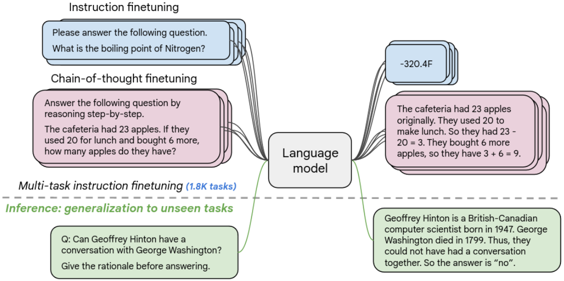
- trained on both encoding and decoding objectives
- also known as text-to-text models
- any task is cast as converting one text to another
- encoding the input text and then decoding the output text
- most popular example: T5 [Raffel et al., 2020]



Instruction fine-tuning and alignment

Helpful instructions

- One can further fine-tune a generative language model on a collection of specific datasets phrased as instructions (check out open FLAN-T5 family of models [Chung et al., 2022])
- sort of an extension of the text-to-text idea
- shown to generalize on unseen tasks
- of course, manually annotated datasets are required.



-	Random	25.0
-	Average human rater	34.5
May 2020	GPT-3 5-shot	43.9
Mar. 2022	Chinchilla 5-shot	67.6
Apr. 2022	PaLM 5-shot	69.3
Oct. 2022	**Flan-PaLM 5-shot**	**72.2**
	Flan-PaLM 5-shot: CoT + SC	**75.2**
-	Average human expert	89.8
	Jun. 2023 forecast (Hypermind)	73.2
	Jun. 2024 forecast (Hypermind)	75.0
	Jun. 2023 forecast (Metaculus)	82.7
	Jun. 2024 forecast (Metaculus)	87.6

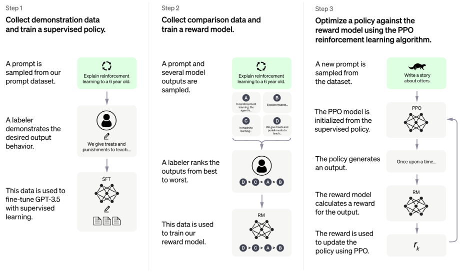
Human-in-the-loop

Important addition: large-scale human supervision (a.k.a. RLHF).

Human-in-the-loop

Important addition: large-scale human supervision (a.k.a. RLHF).

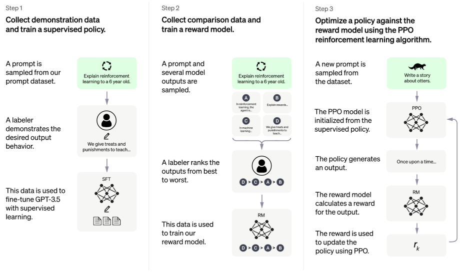
- InstructGPT model [Ouyang et al., 2022]
- pre-trained LM is additionally refined on human preferences: reinforcement learning with human feedback (RLHF)
- human supervision on hundreds of thousands of interactions (crowd-workers paid 2$/hour max)
- pushes the models towards being helpful, harmless, and honest in chat
- often called 'alignment': this very case when an external signal is required, beyond pure language modeling [Rafailov et al., 2023]



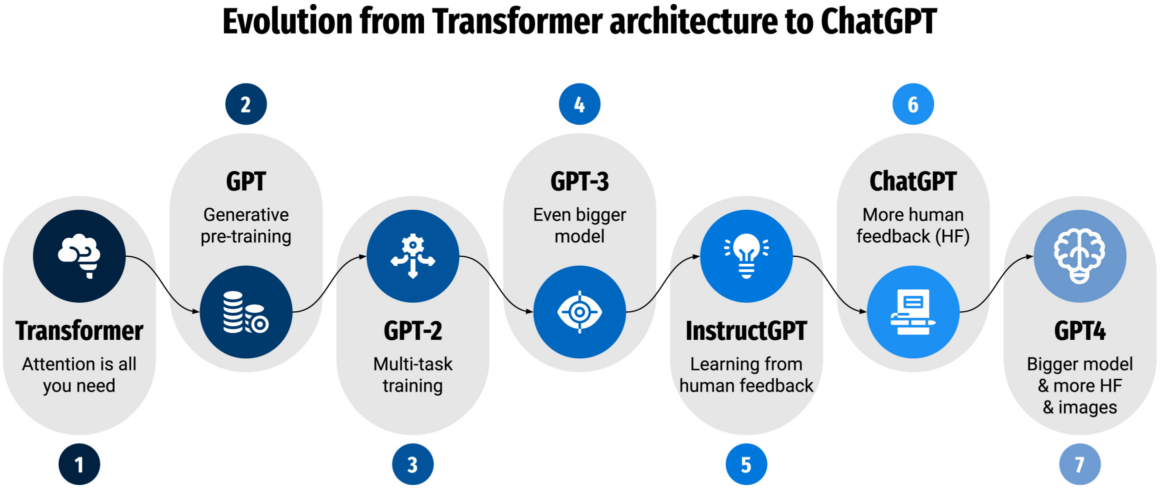
Human-in-the-loop

Important addition: large-scale human supervision (a.k.a. RLHF).

- InstructGPT model [Ouyang et al., 2022]
- pre-trained LM is additionally refined on human preferences: reinforcement learning with human feedback (RLHF)
- human supervision on hundreds of thousands of interactions (crowd-workers paid 2$/hour max)
- pushes the models towards being helpful, harmless, and honest in chat
- often called 'alignment': this very case when an external signal is required, beyond pure language modeling [Rafailov et al., 2023]

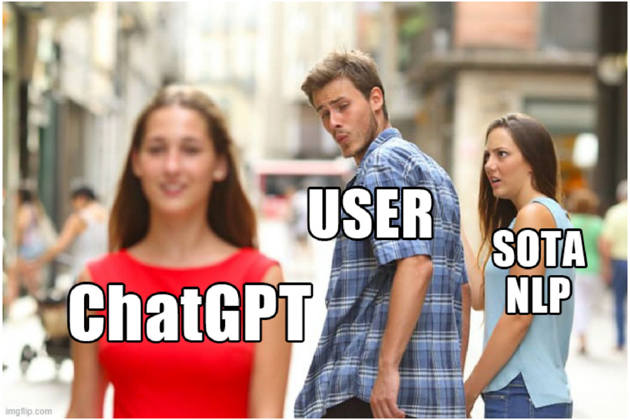


Some even suggest to call such LMs 'instruction-tuned text generators' [Liesenfeld et al., 2023]

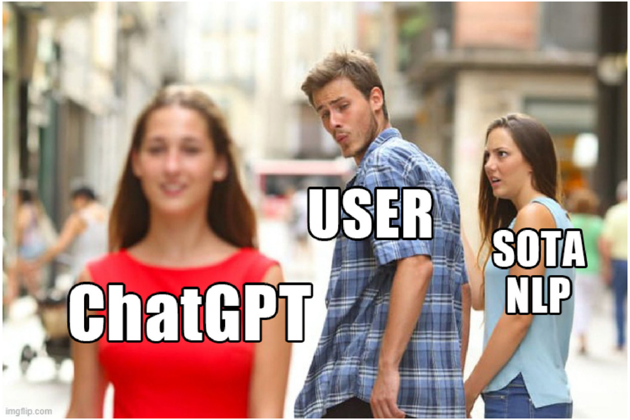


[Kocoń et al., 2023]

LLMs are doing well

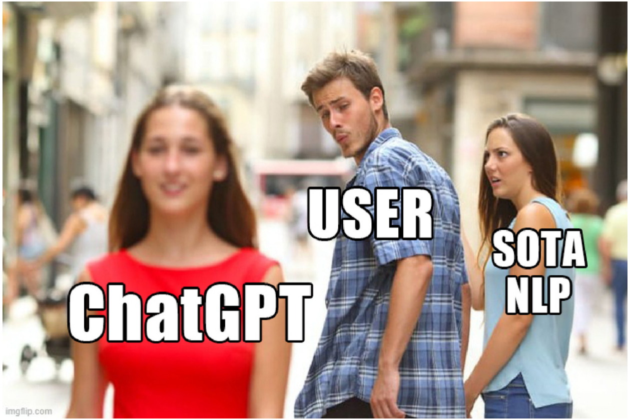


LLMs are doing well



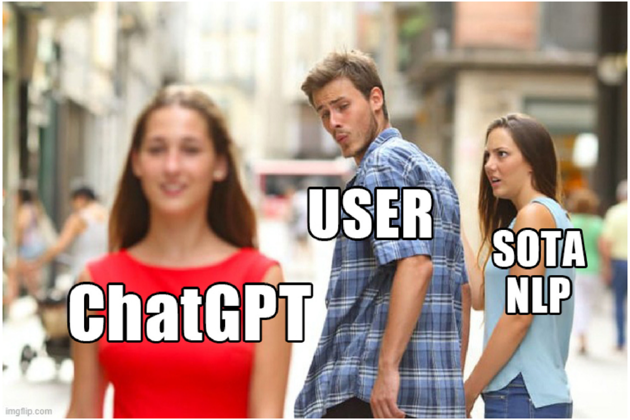
- ChatGPT is not very novel scientifically, but it is a gem of engineering and marketing.

LLMs are doing well



- ChatGPT is not very novel scientifically, but it is a gem of engineering and marketing.
- ChatGPT/GPT-4 are not the superior LMs; they did not destroy NLP
- Large generative LMs are not bad in linguistic tasks, but what does it bring us to?

LLMs are doing well



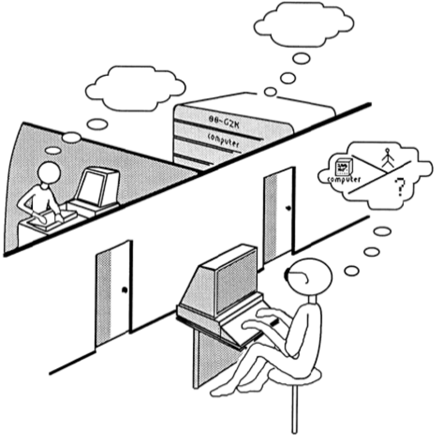
- ChatGPT is not very novel scientifically, but it is a gem of engineering and marketing.
- ChatGPT/GPT-4 are not the superior LMs; they did not destroy NLP
- Large generative LMs are not bad in linguistic tasks, but what does it bring us to?

Contents

Can LLMs have intentions or agency?

- A modern take on the Turing test:
- perceived intelligence (or agency) lies in the eye of the beholder:
- claims of intelligence/agency are meaningful only when their evaluator is taken into account

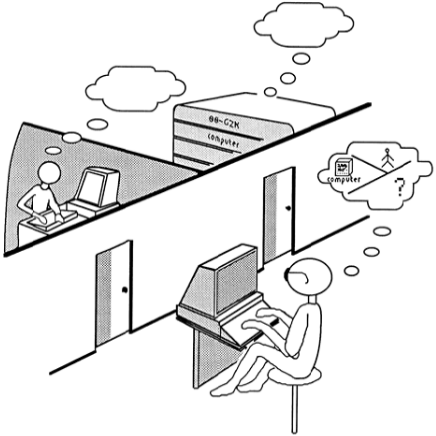
  [Murty et al., 2023]



Can LLMs have intentions or agency?

- A modern take on the Turing test:
- perceived intelligence (or agency) lies in the eye of the beholder:
- claims of intelligence/agency are meaningful only when their evaluator is taken into account

  [Murty et al., 2023]



Still, the answer of my beholder is clear 'no'. Here's why.

Intelligent octopuses and Chinese rooms

- Language skills $\neq$ intelligence or agency [Bender and Koller, 2020]

Intelligent octopuses and Chinese rooms

- Language skills $\neq$ intelligence or agency [Bender and Koller, 2020]
- Humans do use language as a substrate for knowledge [Mahowald et al., 2024]...
  - that's why some functional skills can be learned from texts
  - even passing some form of the Turing test...

Intelligent octopuses and Chinese rooms

- Language skills ≠ intelligence or agency [Bender and Koller, 2020]
- Humans do use language as a substrate for knowledge [Mahowald et al., 2024]...
  - that's why some functional skills can be learned from texts
  - even passing some form of the Turing test...
- ...even if all of this is done by a 100% automaton.

Can LLMs have intentions or agency?

What is this?

Can LLMs have intentions or agency?

What is this?

Float pool, also known as sensory deprivation tank or isolation tank.

Image source: Wikipedia

Can LLMs have intentions or agency?

Lack of permanent awareness/processing

- LLM frameworks are executable computer code by design
- they only respond to stimuli (prompts)
- when no prompt is given, LLM ‘is not running’:
  - no ‘contemplation’ or ‘thinking over’ or ‘making decisions’
- as any computer program, they stop when they reach the end of the code/function.

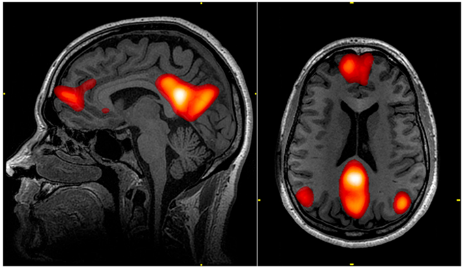
Can LLMs have intentions or agency?

Lack of permanent awareness/processing

- LLM frameworks are executable computer code by design
- they only respond to stimuli (prompts)
- when no prompt is given, LLM 'is not running':
  - no 'contemplation' or 'thinking over' or 'making decisions'
- as any computer program, they stop when they reach the end of the code/function.

Unlike us humans!

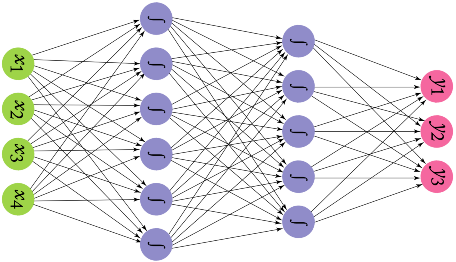
- Default state network in the brain 'integrates meaning over long period of time' [Buckner and DiNicola, 2019]
- Humans 'contemplate' even without any external stimuli
  - e.g., in a sensory deprivation tank
- humans are always 'online'
- I believe this is a *sine qua non* for agency.

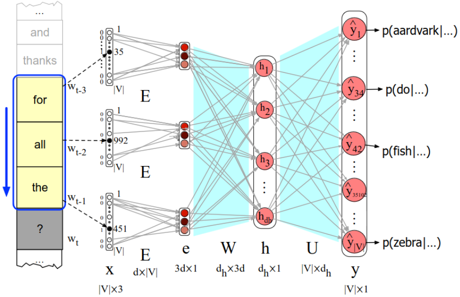


Can LLMs have intentions or agency?

No substance for agency

- LLMs are sets of numerical weights in a large multi-nomial classifier
- basically, a bunch of matrices (tables) with float numbers
- ...and a few rules on converting natural language words into vectors and multiplying them by the matrices
- What exactly can be an agent here?

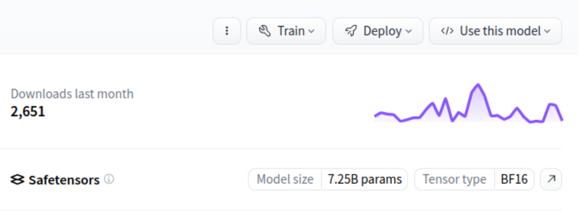




Can LLMs have intentions or agency?

‘Digital’ means ‘easy to copy’

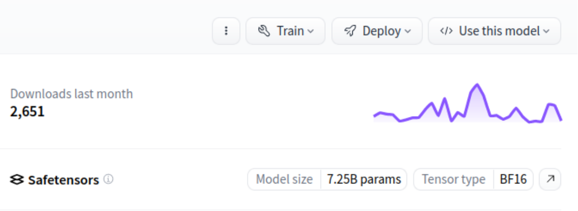
- Technically, any number of absolutely identical copies of any LLM can be created any time.
- Will they all have the same ‘intentions’?
- Will they all be one and the same ‘agent’?
- Looks very ill-defined to me.



Can LLMs have intentions or agency?

‘Digital’ means ‘easy to copy’

- Technically, any number of absolutely identical copies of any LLM can be created any time.
- Will they all have the same ‘intentions’?
- Will they all be one and the same ‘agent’?
- Looks very ill-defined to me.



Widespread ‘anthropomorphisation’ of LLMs can be partially caused by the influence of commercial closed-source models: one cannot download or copy them.

Contents

Questions and answers

Agency? Intentions?

- Language models (LMs) estimate probabilities of linguistic sequences...
- ...but in addition, they can be used directly for text generation (chat-bots).
- Generative LMs are becoming a significant part of our lives

Questions and answers

Agency? Intentions?

- Language models (LMs) estimate probabilities of linguistic sequences...
- ...but in addition, they can be used directly for text generation (chat-bots).
- Generative LMs are becoming a significant part of our lives
- Modern large LMs based on deep artificial neural networks are much better than LMs of the past in capturing linguistic structure (at least for English).
- But they are not anything like 'agents', and they can't have 'intentions'.
- They are 'libraries, not librarians', despite the opinion in [Lederman and Mahowald, 2024]:

Questions and answers

Agency? Intentions?

- Language models (LMs) estimate probabilities of linguistic sequences...
- ...but in addition, they can be used directly for text generation (chat-bots).
- Generative LMs are becoming a significant part of our lives
- Modern large LMs based on deep artificial neural networks are much better than LMs of the past in capturing linguistic structure (at least for English).
- But they are not anything like 'agents', and they can't have 'intentions'.
- They are 'libraries, not librarians', despite the opinion in [Lederman and Mahowald, 2024]:
  - for many reasons, including the inherent lack of default state system.
- LLMs are only machines trained to reproduce the probability distribution for the next words given the previous lexical context.

References I

Bender, E. M. and Koller, A. (2020).
Climbing towards NLU: On meaning, form, and understanding in the age of data.
In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5185–5198, Online. Association for Computational Linguistics.

Bengio, Y., Ducharme, R., and Vincent, P. (2003).
A neural probabilistic language model.
Journal of Machine Learning Research, 3:1137–1155.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020).
Language models are few-shot learners.

References II

- Buckner, R. L. and DiNicola, L. M. (2019).
  The brain's default network: updated anatomy, physiology and evolving insights.
  Nature Reviews Neuroscience, 20(10):593–608.

- Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al. (2022).
  Scaling instruction-finetuned language models.
  arXiv preprint arXiv:2210.11416.

- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019).
  BERT: Pre-training of deep bidirectional transformers for language understanding.
  In Burstein, J., Doran, C., and Solorio, T., editors, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

References III

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2023).
Mistral 7b.

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., Kocoń, A., Koptyra, B., Mieleszczenko-Kowszewicz, W., Miłkowski, P., Oleksy, M., Piasecki, M., Łukasz Radliński, Wojtasik, K., Woźniak, S., and Kazienko, P. (2023).
ChatGPT: Jack of all trades, master of none.
Information Fusion, 99:101861.

Lederman, H. and Mahowald, K. (2024).
Are language models more like libraries or like librarians? bibliotechnism, the novel reference problem, and the attitudes of llms.

References IV

Liesenfeld, A., Lopez, A., and Dingemanse, M. (2023).
Opening up chatgpt: Tracking openness, transparency, and accountability in instruction-tuned text generators.
In Proceedings of the 5th International Conference on Conversational User Interfaces, pages 1–6.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. (2024).
Dissociating language and thought in large language models.
Trends in Cognitive Sciences.

Murty, S., Paradise, O., and Sharma, P. (2023).
Pseudointelligence: A unifying lens on language model evaluation.
In Bouamor, H., Pino, J., and Bali, K., editors, Findings of the Association for Computational Linguistics: EMNLP 2023, pages 7284–7290, Singapore. Association for Computational Linguistics.

References V

- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022).
Training language models to follow instructions with human feedback.
arXiv preprint arXiv:2203.02155.

- Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. (2023).
Direct preference optimization: Your language model is secretly a reward model.
In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, Advances in Neural Information Processing Systems, volume 36, pages 53728–53741. Curran Associates, Inc.

- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J., et al. (2020).
Exploring the limits of transfer learning with a unified text-to-text transformer.
J. Mach. Learn. Res., 21(140):1–67.

References VI

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. (2021).
The neural architecture of language: Integrative modeling converges on predictive processing.
Proceedings of the National Academy of Sciences, 118(45):e2105646118.

Shannon, C. E. (1948).
A mathematical theory of communication.
The Bell system technical journal, 27(3):379–423.