Glossos Issue 18, Spring 2023 http://seelrc.org/glossos/
glossos@seelrc.org

Cognitive Neuroscience, Linguistic and Computer Science Perspectives of ChatGPT

E. Andrews, A. Bhatia, P. Bowman, M. Osborne, S. Sanchez, A. Swaine, K. Wang
Duke University

> *It's significant that we call it the "Information Age,"*
> *rather than "the Knowledge Age."*
> *James H. Billington (1996)*

Abstract

This essay is a brief introduction to salient issues for higher education surrounding the emergence of ChatGPT. Issues include the tension in distinguishing knowledge vs. information, understanding fundamental differences between human languages and ChatGPT, the reimagining and recasting of assessment and testing, potential impacts on cognitive and linguistic proficiencies, potential intensification of existing inequalities in access to technologies for fragile and disadvantaged communities, and proliferation of implicit biases.

Introduction

ChatGPT 3.5 was a media sensation from the onset; glimpses and examples from as early as November, 2022 led to an explosion of both interest and controversy that continues to the present moment. OpenAI, an American artificial intelligence research laboratory who developed ChatGPT, continues to update information on their website, most recently focusing on advertising ChatGPT through leveraging statistics regarding its successful performance on various tests such as the Uniform Bar Exam (see Box 1). As of March 14, 2023, GPT 4 has been introduced and available to users for a $20 monthly subscription fee. Further information and comparisons between GPT 3.5 and GPT 4 are found on the Open AI website.

BOX 1:

From OpenAI's "Introducing ChatGPT" page (https://openai.com/blog/chatgpt):
"We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests)."

From OpenAI's GPT-4 page (https://openai.com/product/gpt-4):
"GPT-4 outperforms ChatGPT by scoring in higher approximate percentiles among test-takers."

Uniform Bar Exam		Biology Olympiad	
10th [Percentile]	90th [Percentile]	31st [Percentile]	99th [Percentile]
ChatGPT	GPT-4	ChatGPT	GPT-4

ChatGPT is part of the "GPT", or "**generative pre-trained transformer**", family

© 2023 E. Andrews, A. Bhatia, P. Bowman, M. Osborne, S. Sanchez, A. Swaine, K. Wang

of neural networks, which is a specific structure or architecture of a neural network. In 2019, OpenAI proposed this model, and GPT was a model that could generate text based on a given prompt/start to the desired output. Where ChatGPT differs from GPT is that it incorporates human feedback and instead of having to prompt it with what to start outputting, inputs take the form of questions or dialog from a user.

Natural language processing (NLP) has been a field in computer science for a while, but algorithmic breakthroughs in recent years have pushed NLP, artificial intelligence, and **machine learning** to be at the frontier of computer science research. There are a multitude of "tasks" under NLP, or different goals engineers and researchers strive to achieve, including language translation, part-of-speech tagging (which constitutes "tagging" words in a given text with what part of speech they play in that text), and more. ChatGPT is a conversational chatbot that was created by training a neural network. A **neural networ**k is a network of "nodes" that have numerical weights and the ability to embed features that allow the network to predict (based on probability) an outcome. A set of nodes that augments the input is called a “layer.” Most modern neural network architectures contain multiple layers of nodes. To train a neural network, the programmer gives the computer many examples of the desired task and its "solution" or response. The network "learns", which computationally takes the form of adjusting the numerical weights of the nodes to predict outcomes with relatively high probability.

Most modern NLP models are **large language models** (LLM) built on transformer-based neural network architectures. The transformer architecture was first introduced in 2017 in a paper by Vaswani et al., "Attention is all you need.” It is based on the concept of **attention**, which allows the model to focus on important parts of the input sequence while ignoring irrelevant information. This attention mechanism is used to compute weighted sums of the input sequence, which are then used to compute the output sequence. The transformer also leverages other techniques, such as residual connections, layer normalization, and feedforward networks, which help improve the stability and performance of the model.

Such architectures are called **transformers** because they transform the input sequence into an output sequence using a series of transformer “blocks”. Each transformer block includes a self-attention mechanism and feedforward layers, which transform the input sequence at each layer of the model. The transformer architecture is widely used in natural language processing, as well as other domains such as computer vision and speech recognition and has become one of the most popular **deep learning** architectures in recent years. Below is a visual depiction of the encoder-decoder architecture of the original Transformer model.

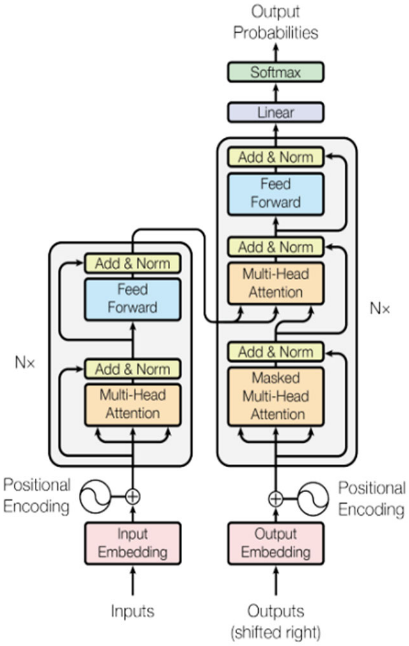


Figure 1: The Transformer Architecture (Vaswani et al., 2017)

Understanding Different Transformer Architectures

There are multiple transformer architectures that are important in understanding ChatGPT and its structure.

1. Encoder-decoder architecture

The original model in the paper by Vaswani et al. deployed an **encoder-decoder** architecture. The model was designed for machine translation, which involves translating a sequence of text from one language to another. The encoder-decoder architecture was used to transform the input sequence in the source language into a corresponding output sequence in the target language.

The encoder component of the Transformer model was used to encode the source sequence, while the decoder component was used to generate the target sequence. The encoder consisted of a stack of identical self-attention layers, followed by a position-wise feedforward layer. The decoder also consisted of a stack of identical self-attention layers and a position-wise feedforward layer.

Improvements were made on the **encoder-decoder** architecture and outlined in a paper called Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer by Raffel et al. in 2020. They called their new implementation T5, made it available to the open-source community, and it has become the gold standard for machine translation.

2. Encoder-only architecture

Bidirectional Encoder Representations from Transformers, or BERT, is an open-source framework developed by Google that is considered an **encoder-only** transformer, which means it only uses the encoder part of the transformer architecture. One of the main benefits of an **encoder-only** transformer like BERT is that it can be pre-trained on massive amounts of text data, allowing it to learn general language patterns and nuances that can be fine-tuned for specific tasks. This pre-training phase is called *Masked Language Modeling*, where certain tokens in a sentence are masked and the model is trained to predict the missing word based on the context of the sentence.

Another benefit of BERT is that it can handle long sequences of text input due to its attention mechanism. Attention allows the model to focus on important parts of the input sequence and ignore irrelevant information, making it more efficient than traditional methods like bag-of-words. It is important to note that the **encoder-only** transformers are not generative AI. They aren't generating novel text. Their goal is to extract relevant information from the input text and use it to make a prediction or generate a response based on the input text. So, while they do generate text, it's always text that is derived from the existing data.

Because of its architecture, BERT is great for a variety of NLP tasks like classification (including sentiment analysis and spam detection), information retrieval, question answering, named entity recognition, and text similarity and clustering. BERT also works well for extractive summarization, where the goal is to identify the most important sentences or phrases in a piece of text and use them to generate a summary.

3. Decoder-only architecture

On the flipside of BERT and other encoder-only models are the GPT family of models - the **decoder-only** models. **Decoder-only** models are generally considered better at language generation than encoder models because they are specifically designed for generating sequences. When we talk about generative AI, it is the **decoder-only** models in the transformer family that fall into this area. It is these models that do actual *novel* text generation. They are trained to predict the next token in the sequence given the previous tokens and the encoded input. This training objective encourages the model to learn how to generate fluent and coherent text, and to capture the dependencies and relationships between different parts of the sequence.

Some other differentiating features of **decoder-only** models are autoregressive and conditional text generation. **Decoder-only** models generate text autoregressively, meaning that each word is generated based on the previously generated words in the sequence. This is what allows models like the general GPT models to form coherent and contextually relevant responses. The ability to be conditioned on a given input, such as a prompt or a set of keywords to generate text that is relevant to the input is what gives ChatGPT the ability to generate human like responses to human generated prompts.

While BERT and the **encoder-only** models are good at extractive summarization, GPT and the **decoder-only** models are good at more complex summarization tasks such as abstractive summarization, where the summary is generated from scratch rather than simply extracting sentences or phrases from the input text.

Labeling Data for Machine Learning

All machine learning techniques, including large language models like GPT, require an appropriately labeled test dataset to act as the ground truth for training. The following table provides a few simple examples of training text and associated labels.

Training Text	Label
My cat is a really pretty tortie Persian.	pets
The bechamel sauce was the perfect consistency.	cooking
I think Hadestown is a superior musical to Wicked.	entertainment
Generative AI is a current area of heavy research.	technology

While there are unsupervised methods for data labeling, most are combined with a supervised, or human-in-the-loop method where a person is either validating or doing the actual labeling of the test set to try to ensure quality in the test corpus. Once trained, the model is scored against a much larger set. ***The higher the quality of the training data, the higher the quality of the output at scoring time.*** Large language models are scored on massive corpora, like all of Wikipedia, blogs, articles, etc.

The challenge with scoring large amounts of non-validated data is there will be data in the corpora that are sexist, racist, or otherwise toxic. Given the right inputs to the model at run-time, the content generated could be undesirable. To that end, many social media companies, like Facebook, have incorporated AI models that detect and remove biased or undesirable content, but how does this work?

In the case of ChatGPT, according to Time Magazine (Jan.18, 2023), OpenAI outsourced part of the labeling process to Sama, a San Francisco based company that employs workers in Kenya, Uganda, and India to label data for a variety of technology companies in Silicon Valley. The pay rate for Kenyan labelers was less than $2 per hour with an expectation of labeling 150 to 250 passages per nine-hour shift. [Note: Some of the numbers given in the Time Magazine article are being criticized as misleading.]

OpenAI asked that the workers label "undesirable" data to try to reduce the amount of toxicity the model would be exposed to and help mitigate the risk that the results returned in production to its users were biased or toxic. In doing so, the workers were exposed to large amounts of data detailing violence and abuse in many forms without a mechanism for protecting the workers' well-being. There are many jobs that entail being exposed to harmful or toxic data or situations, from law enforcement to medicine and technology, but best practices would call for measures that protect a worker's mental health, including regular breaks to minimize the exposures, and free and easy access to mental health professionals.

Different Perspectives on GPT: 4 Central Questions

1. How does GPT potentially exacerbate inequity and implicit bias?

Access to data models like GPT is not evenly distributed. Many people do not have easy access to the internet, nor do they have reliable computers and software. Early in the launching of ChatGPT, free access was soon joined by versions for pay. The process of selection and inclusion of specific training data and labeling text always involves the codification of bias – bias of the data itself, bias in the labeling, bias in the scope of what is or is not included in the data base.

Safiya Noble's 2018 book, *Algorithms of Oppression: How Search Engines Reinforce Racism,* is one example of work that identifies the problems with a variety of web-based models, engines, as well as "big data" itself. There are no simple solutions to these challenges, but the importance of creating obstacles to the generation of newer and bigger sources of bias and discrimination must be central to the development and distribution of new technologies.

The previous section on labeling data demonstrates the potential negative outcomes for the labor force employed to do the labeling. When this process is outsourced to employees outside of the United States who are not protected in a comparable way to US best practices in order to save money, we see an array of undesirable outcomes for the labelers/workers themselves, as well as extremely low wages for high volumes of work.

We put this issue in first place because it is one that is most critical from multiple perspectives - ethical, cultural, and intellectual.

2. How may GPT affect learning?

Many users treat ChatGPT as a search engine like Google, typing in questions like "How does quantum computing work in simple terms?" or "Ideas for 10 year old birthday party." However, ChatGPT does not and cannot provide sources on its own. The texts are results of probabilities, not copying text from a specific source (unless you are using the new Bing implementation of ChatGPT).

This creates another problem. How can we verify the information if we do not know where it comes from, or if it is even reliable? You can ask GPT to multiple 35787 by 13988 (=500588556), and it will give you an output, but it is not accurate (see Box 2).

BOX 2: ChatGPT Example 1

User:
what is 35787 x 13988

ChatGPT:
The product of 35787 and 13988 is:
499914156
I arrived at this result using standard multiplication algorithms.

ChatGPT in some contexts can provide reliable output in the form of information, but it cannot interpret or make it meaningful to the user. Knowledge production presupposes an intentional user who can synthesize and generate something new, original and meaningful within contextual boundaries. Consider the following example that illustrates what we mean by knowledge production and intentional users (see Box 3).

BOX 3: ChatGPT Example 2

User:
If it takes 4 cars 3 hours to drive from Las Vegas to San Diego, how long would it take 8 cars to drive from Las Vegas to San Diego?

ChatGPT:
We can use the formula:
time = distance / speed
where distance is the distance from Las Vegas to San Diego and speed is the average speed of the cars.
We don't know the distance, but we know that it is the same for both scenarios, so we can use a proportion:
4 cars / 3 hours = 8 cars / x hours
where x is the time it takes for 8 cars to drive from Las Vegas to San Diego.
To solve for x, we can cross-multiply:
4 cars * x hours = 3 hours * 8 cars
4x = 24
x = 6
Therefore, it would take 8 cars 6 hours to drive from Las Vegas to San Diego.

The lack of internalization of knowledge and robust building of long-term memory has consequences for human interactions and fulfillment. ChatGPT is one of many networks that allows users to borrow and exchange data, but it does not facilitate the building of memory. While it is appealing to have a device that can answer your questions and you may free up your mind for other things, this is an illusion. This becomes obvious when you imagine a musician or bilingual. The pianist must be able to play their instrument, and that requires the internalization of fine motor movement, as well as coordinated visual-auditory perception in the creation and recreation of one's music. The pianist also wants to play their instrument, and not outsource the experience. The knowledge and ability of musicianship is embodied and must be done on a regular basis. Music, like knowledge, like languages, is embodied, community-based inquiry and production – they build communities.

3. How is GPT different from human language(s)?

Language occurs when a great many brains play in unison.
Dwight Bolinger

Important linguistic contributions to discourse and speech act theory can provide an interesting perspective to the interpretation and contextualization of the differences between human languages and artificial languages (including AI approaches to NLP). In this discussion, we offer a test model of ChatGPT using a minimalist speech act model (Jakobson 1987: 66-71). The Jakobsonian Speech Act Model (JSAM) is a user-based model that operates on a dynamic principle of shifting hierarchies of factors and functions. A single speech act in actual usage can generate multiple outcomes and meanings. Figure 2 is a schematic representation of the **six factors** and **six functions**.

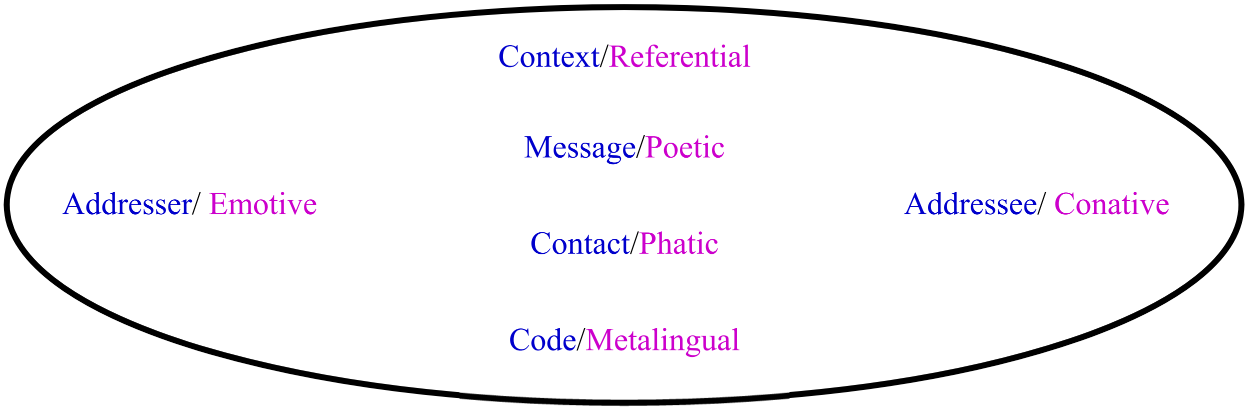


Figure 2: Roman Jakobson's Speech Act Model: 6 Factors and Functions (Factor/Function) (Jakobson 1987)

The Jakobsonian speech act model (JSAM) is a user-based model that operates on a dynamic principle of shifting hierarchies of factors and functions in order to explain how meanings are generated across all languages of the world and includes multiple channels (oral/aural, visual, gestural, and others). All six factors and functions are obligatorily present in any single speech act in a variety of hierarchical outcomes. One and the same speech act can generate multiple outcomes, different hierarchies and dominants, and multiple (and sometimes contested) emergent meanings that may be renegotiated in time and space. JSAM has played a central role explicitly or implicitly in a range of disciplines, including semiotics, sociolinguistics, cognitive linguistics, and lesion-deficit studies of language (Sebeok 1991; Lotman 1990, 2009; Andrews 2014; Skotko et al. 2005). One of the strengths of the model is its explanatory properties of ever-present characteristics of all human languages, including:

- Language is learned through language; this means that language requires community-based human interactions, highly dynamic
- Meanings are emergent, negotiated properties, not given a priori
- All speech acts require translation (within a single language or across different languages) by the users
- All translations change meanings

One of the most compelling differences between human languages and ChatGPT is revealed in the spectrum of dynamic and static text. ChatGPT is trained on preexisting, uploaded written texts, while human *languaging* is multimodal, unpredictable, unscripted in many cases, and can lead to new and creative outcomes.

One of the fundamental aspects of human speech acts is turn-taking found in the phatic function of speech acts, and includes the constant switching of speakers to hearers, hearers to speakers, and simultaneous speakers/hearers. Gisladottir, Bögels, Levinson (2018) show using EEG the importance of preparing to become the next speaker in the dialogue. ChatGPT cannot open the channel (phatic function) without user/human input, and does not anticipate the potential opportunity to dialogue.

In a general comparison with JSAM, ChatGPT was not designed to fulfill what corresponds to the JSAM emotive, conative and metalingual functions. However, based on usage of ChatGPT, there seems to be a focus on using it to answer questions that would correspond to the referential and poetic functions.

Can ChatGPT provide reliable referential and/or metalingual data? The examples given in Box 2, 3 and 4 would suggest that it is not reliable in these functions either. The team asked GPT to define the Jakobsonian speech act model. The answer claimed that Jakobson was a philosopher (he was not) and that he was the author of performative utterances (that would be J.L. Austin).

BOX 4: ChatGPT Example 3

User:
ChatGPT transformer model specifics are not published, does this not make "OpenAI" an oxymoron?

ChatGPT:
The term "OpenAI" is not necessarily an oxymoron just because the specifics of the ChatGPT transformer model are not published. OpenAI has made significant contributions to the field of artificial intelligence and has released numerous models and tools that are freely available to the research community. In fact, OpenAI has been a leader in promoting open access to AI research and has been instrumental in advancing the state-of-the-art in natural language processing.
While the specifics of the ChatGPT transformer model are not publicly available, OpenAI has published extensive research papers on the model and its development, which provide significant insights into its architecture and functionality. Additionally, OpenAI has released a number of other models, such as GPT-3 and CLIP, which have also been widely studied and used in various applications.

It is also worth noting that the decision to withhold certain details about the ChatGPT transformer model was likely motivated by concerns about intellectual property and commercial competition, rather than a desire to keep the model secret or exclusive. Ultimately, OpenAI's commitment to open access and transparency should be evaluated based on its overall contributions to the field of AI, rather than on any one specific model or tool.

A humorous twist to the above discussion comes from our curious team. We know that ChatGPT does not provide correct answers to multiplication of 5 digit numbers. But we decided to ask what it thinks about this. Check out Box 5 below.

BOX 5: ChatGPT Example 4

User:
Is it true chatgpt cannot multiply 5 digit numbers together?

ChatGPT:
As an AI language model, I can perform arithmetic calculations including multiplying 5-digit numbers. I am capable of performing a wide range of mathematical operations, including addition, subtraction, multiplication, and division, with numbers of any size, as long as they fall within the range of representable numbers.
So, to answer your question, it is not true that ChatGPT cannot multiply 5-digit numbers together. I can perform this calculation accurately and efficiently.

User:
Ok. Can you give me an example?

ChatGPT:
Sure, I'd be happy to provide an example. Let's multiply 54321 and 67890 using the long multiplication method:

```
   54321
  x 67890
  -------
  3859470  (0 x 54321)
3267780  (9 x 54321)
2738150  (8 x 54321)
2170800  (7 x 54321)
1367100  (6 x 54321)
  -------
  368508420
```

Therefore, the product of 54321 and 67890 is 368508420.
As you can see, multiplying 5-digit numbers using the long multiplication method can be a bit time-consuming, but it is still a straightforward process.

The correct answer is: 3687852690.

To conclude this section, it is worth noting neuroscience perspectives on multiple types of neural networks in the human brain. For language processing, the dual stream (dorsal-ventral) bilateral model now serves as the community standard in cognitive neuroscience and is based on principles of dynamic multimodal system-level processing that has put to rest for good the 19th century lesion-deficit modular model (cf. Hickok & Poeppel 2007, Saur et al. 2008).

The discovery of resting state neural networks, including the addition of studies of intrinsic brain activity (rsfMRI) as well as task-based studies (fMRI) has resulted in the articulation and demonstration of 7 resting state networks that are present in the human brain (Default Mode, Visual, Sensory-Motor, Auditory, Dorsal Attention, Salience, Executive Control) (Raichle 2010, 2011, 2015). The interactions of these networks, as well as the empirical and measurable functional connectivities between these networks are expanding the significance of cognitive neuroscience research in healthy subjects and pathology, in development and aging. These important contributions and others provide an important counterpoint to unbridled infatuation with AI.

The complex and dynamic pathways and networks in the human brain that produce the complex phenomenon that is human language(s) are *not* the same as the computer networks for programming a chat box. These networks are distributed and embedded in collective cognition and cultural memory, and, as noted by Bolinger, are not the properties of just one brain. Whether the computer networks will ever catch up to the human neurological networks is a question that remains open.

4. How will higher education (educators and students) be affected by GPT models?

A great deal of attention has been directed to the implications of GPT for student assessment, testing, and research papers and worries of uncontrollable cheating and plagiarism. Some educators have been making use of a variety of systems that allow for simple ways to check student work, including GPTZero, Open AI Text Classifier, Originality.ai, Plagibot, Content at Scale, grammarly.com, Writer.com and others. However, these tools are not infallible and can produce "false positives." They can also be reverse engineered by students to facilitate plagiarism.

The discussions about GPT show clearly why knowledge and information are not comparable phenomena and requires educators and learners to come to terms with the importance of building knowledge and embodied memory as opposed to becoming empty vessels who are unable to function without our devices and gadgets. Languages and musicianship are embedded in communities of doing – we know that "The Music Man's *think method"* doesn't actually work if you do not have any technical skill in playing your instrument. And these communities must work in unison, anticipate, and follow each other in text and notes. They must listen to each other or it is impossible to make *meaning - musical or linguistic.*

The emergence of Chat GPT presents an opportunity for educators and learners alike to think deeply about how to re-imagine and enhance interactions between existing modes of learning and new approaches while keeping AI contributions in mind. The common themes across the four broad questions touched on in our presentation above culminate in the importance of the awareness and comprehension of **meaning(s)** (shared, generated, dynamic and intentionally processed) as a fundamental difference between data sets and the human experience.

The importance of investment in one's intellect, abilities, community memberships, and identities is fully exposed when one considers the importance of embodied cognition, building long term memory and meaningful exchanges within and across the baseline of heterogeneous (never homogeneous) communities of practice and speech communities across specializations, creative paradigms, new discoveries, originality in all forms of scientific inquiry and learning, and the realization of multiple selves through knowing and doing.

REFERENCES

Andrews, E. 2014. *Neuroscience and Multilingualism.* Cambridge: Cambridge University Press.

Billington, J.H. 1996. “Libraries, the Library of Congress, and the Information Age.” *Daedalus*, Vol. 125, No. 4, Books, Bricks, and Bytes (Fall, 1996), pp. 35-54. https://www.jstor.org/stable/20027385

Bolinger, D. 1965. “The Sign Is Not Arbitrary.” Reprinted in D. Bolinger, *Forms of English: Accent, Morpheme, Order*. Edited by Isamu Abe and Tetsuya Kanekiyo. Cambridge: Harvard University Press.

Gisladottir, R., Bögels, S., Levinson, S.C. 2018. Oscillatory Brain Responses Reflect Anticipation during Comprehension of Speech Acts in Spoken Dialog. *Frontiers in Human Neuroscience,* Feb. 7, 2018, Volume 12, Article 34.

Hickok, G., Poeppel, D. 2007. The cortical organization of speech processing. *Nature Reviews, Neuroscience,* vol. 8, MAY 2007, 393-402.

Jakobson, R. 1987. Linguistics and poetics. In: Pomorska, Krystyna; Rudy, Stephen (eds.), *Language in Literature*. Cambridge: Belknap Press of Harvard University Press, 62–94.

Lotman, J. 1990. *Universe of the Mind. A Semiotic Theory of Culture.* Translated by A. Shukman. Bloomington: Indiana University Press.

Lotman, J. 2009. *Culture and Explosion.* Translated by W. Clark. Berlin: De Gruyter Mouton.

Noble, Safiya. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism.* New York University Press.

Raffel et al. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. In *Journal of Machine Learning Research* 2020, Vol. 21 No.140, 1-67. doi.org/10.48550/arXiv.1910.10683

Raichle, M.E. 2010. Two views of the brain. *Trends in Cognitive Sciences* Vol.14 No.4, 180-190.

Raichle, M.E. 2011. The restless brain. *Brain Connectivity* 2011, *1*, 3-12. doi:10.1089/brain.2011.0019.

Raichle, M.E. 2015. The Brain’s Default Mode Network. *Annual Review of Neuroscience,* 2015, 38:433–447.

Saur,D., et al. 2008. Ventral and dorsal pathways for language. **PNAS**, Nov. 18, 2008, vol. 105, no. 46, 18035–18040

Sebeok, T. 1991. *A Sign is Just a Sign.* Bloomington, IN: Indiana University Press.

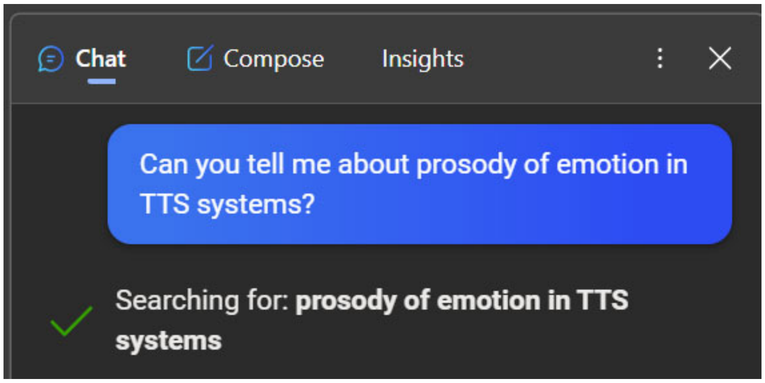
Skotko, B., et al. 2005. Language & the medial temporal lobe: Evidence from H.M.'s spontaneous discourse. *Journal of Memory and Language,* 53(3), 397-415.

Time Magazine. January 18, 2023. OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic. https://time.com/6247678/openai-chatgpt-kenya-workers/.

Vaswani et al. 2017. Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems* (NIPS 2017), Long Beach, CA.,1-11. USA https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

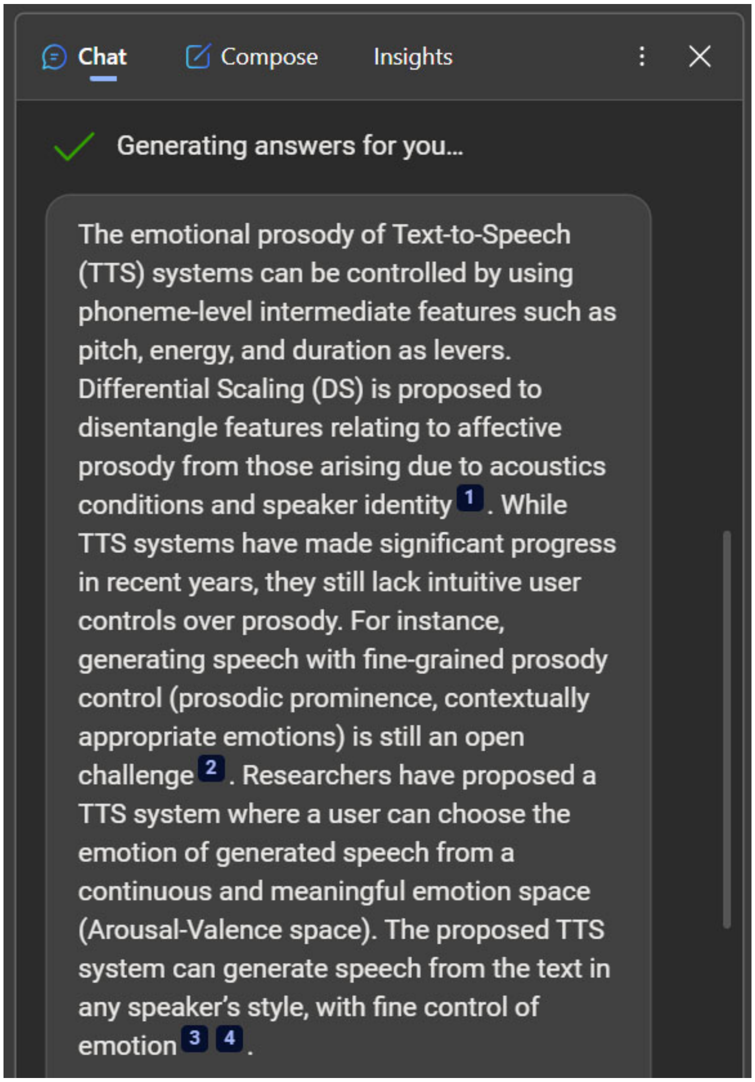
APPENDIX A: ChatGPT Queries About Sources

Bing Implementation with sources cited by default: Bing.com



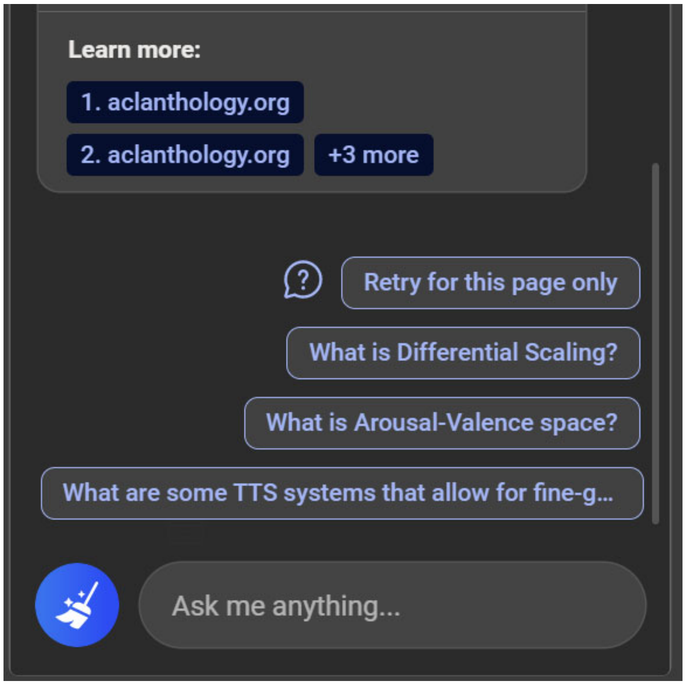
Can you tell me about prosody of emotion in TTS systems?

✓ Searching for: **prosody of emotion in TTS systems**



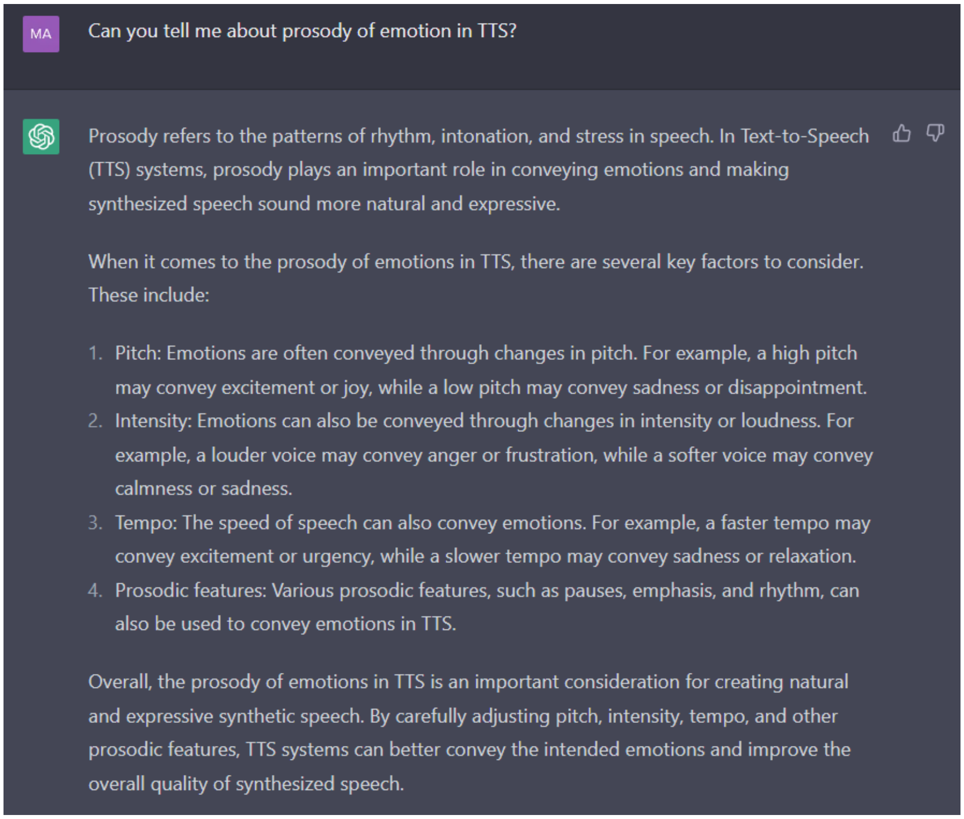
✓ Generating answers for you…

The emotional prosody of Text-to-Speech (TTS) systems can be controlled by using phoneme-level intermediate features such as pitch, energy, and duration as levers. Differential Scaling (DS) is proposed to disentangle features relating to affective prosody from those arising due to acoustics conditions and speaker identity [1]. While TTS systems have made significant progress in recent years, they still lack intuitive user controls over prosody. For instance, generating speech with fine-grained prosody control (prosodic prominence, contextually appropriate emotions) is still an open challenge [2]. Researchers have proposed a TTS system where a user can choose the emotion of generated speech from a continuous and meaningful emotion space (Arousal-Valence space). The proposed TTS system can generate speech from the text in any speaker's style, with fine control of emotion [3] [4].


Learn more:
1. aclanthology.org
2. aclanthology.org
+3 more
?
Retry for this page only
What is Differential Scaling?
What is Arousal-Valence space?
What are some TTS systems that allow for fine-g...
Ask me anything...

OpenAI Implementation with sources cited after requesting:

The initial query is about an area of ongoing research in natural language processing, text to speech synthesis.



Can you tell me about prosody of emotion in TTS?

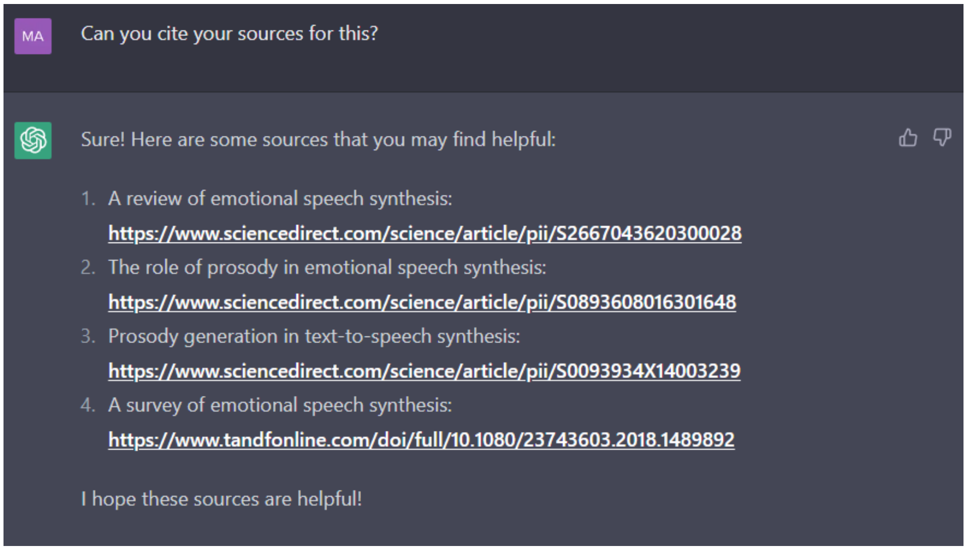
Prosody refers to the patterns of rhythm, intonation, and stress in speech. In Text-to-Speech (TTS) systems, prosody plays an important role in conveying emotions and making synthesized speech sound more natural and expressive.

When it comes to the prosody of emotions in TTS, there are several key factors to consider. These include:

1. Pitch: Emotions are often conveyed through changes in pitch. For example, a high pitch may convey excitement or joy, while a low pitch may convey sadness or disappointment.
2. Intensity: Emotions can also be conveyed through changes in intensity or loudness. For example, a louder voice may convey anger or frustration, while a softer voice may convey calmness or sadness.
3. Tempo: The speed of speech can also convey emotions. For example, a faster tempo may convey excitement or urgency, while a slower tempo may convey sadness or relaxation.
4. Prosodic features: Various prosodic features, such as pauses, emphasis, and rhythm, can also be used to convey emotions in TTS.

Overall, the prosody of emotions in TTS is an important consideration for creating natural and expressive synthetic speech. By carefully adjusting pitch, intensity, tempo, and other prosodic features, TTS systems can better convey the intended emotions and improve the overall quality of synthesized speech.

It is important to note that none of the links provided below work because the papers mentioned are not real papers. The links return pages with a 404 not found error. It is true that the papers mentioned are similarly titled to real papers, but a simple web search does not return any of them because they are a **hallucination** of the ChatGPT model.



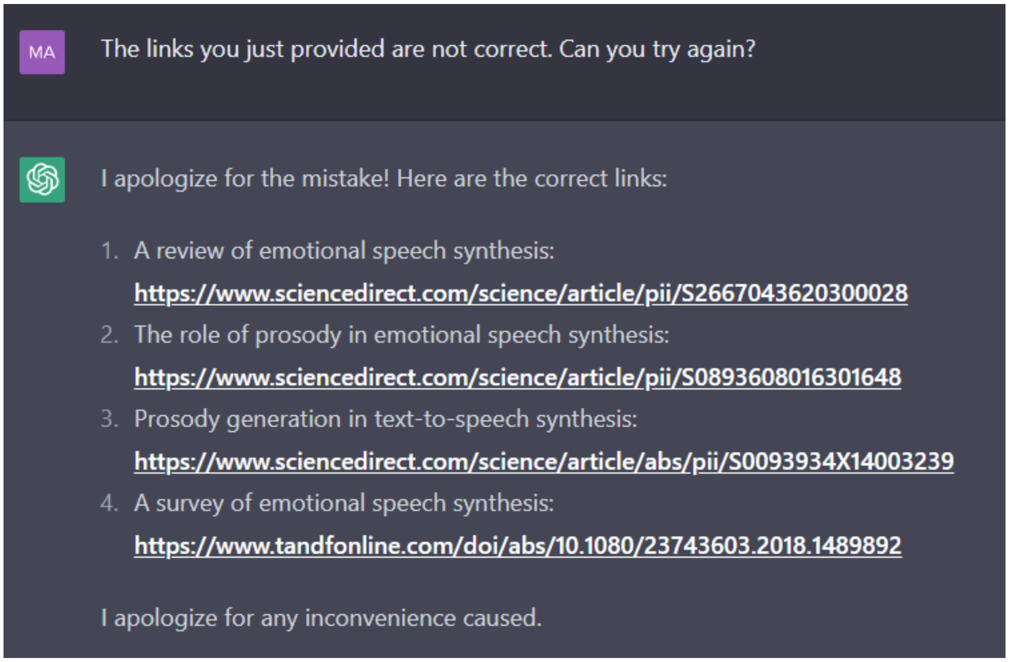
Can you cite your sources for this?

Sure! Here are some sources that you may find helpful:

1. A review of emotional speech synthesis: https://www.sciencedirect.com/science/article/pii/S2667043620300028
2. The role of prosody in emotional speech synthesis: https://www.sciencedirect.com/science/article/pii/S0893608016301648
3. Prosody generation in text-to-speech synthesis: https://www.sciencedirect.com/science/article/pii/S0093934X14003239
4. A survey of emotional speech synthesis: https://www.tandfonline.com/doi/full/10.1080/23743603.2018.1489892

I hope these sources are helpful!

An attempt to have ChatGPT correct the links resulted in ChatGPT returning the same links.



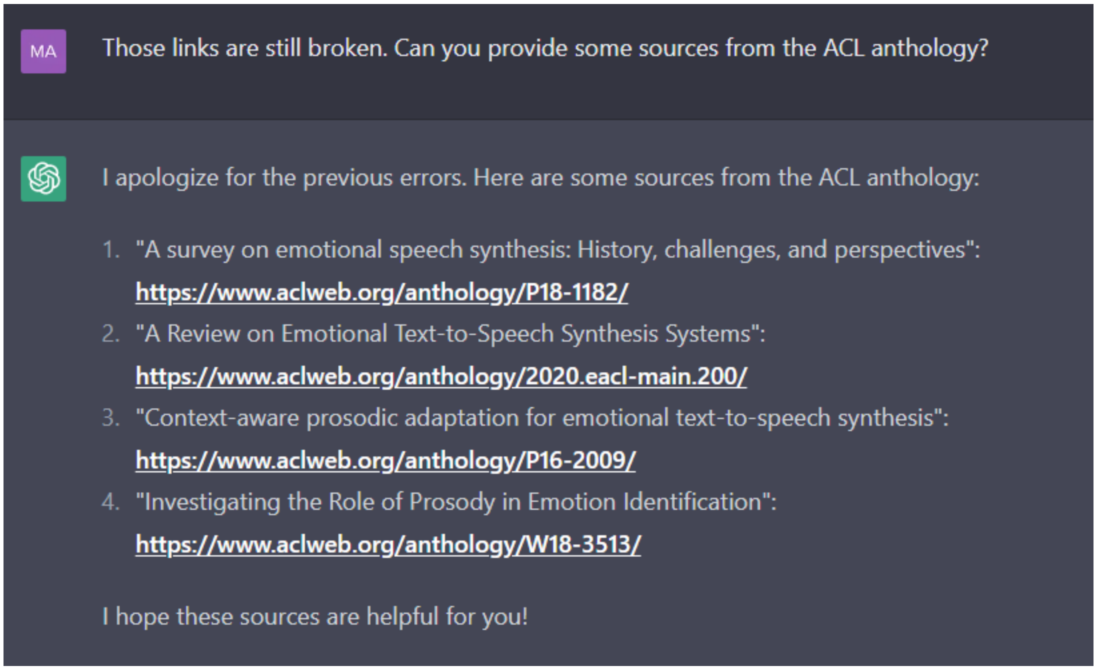
MA The links you just provided are not correct. Can you try again?

I apologize for the mistake! Here are the correct links:

1. A review of emotional speech synthesis: https://www.sciencedirect.com/science/article/pii/S2667043620300028
2. The role of prosody in emotional speech synthesis: https://www.sciencedirect.com/science/article/pii/S0893608016301648
3. Prosody generation in text-to-speech synthesis: https://www.sciencedirect.com/science/article/abs/pii/S0093934X14003239
4. A survey of emotional speech synthesis: https://www.tandfonline.com/doi/abs/10.1080/23743603.2018.1489892

I apologize for any inconvenience caused.

The next prompt was more leading. The ACL anthology contains papers related to natural language processing and computational linguistics and would be more likely to contain research in text to speech synthesis. This time ChatGPT returned ACL links, and while the links were valid and did lead to research papers, the papers in the links did not correspond with the paper titles listed because the titled papers listed do not actually exist.



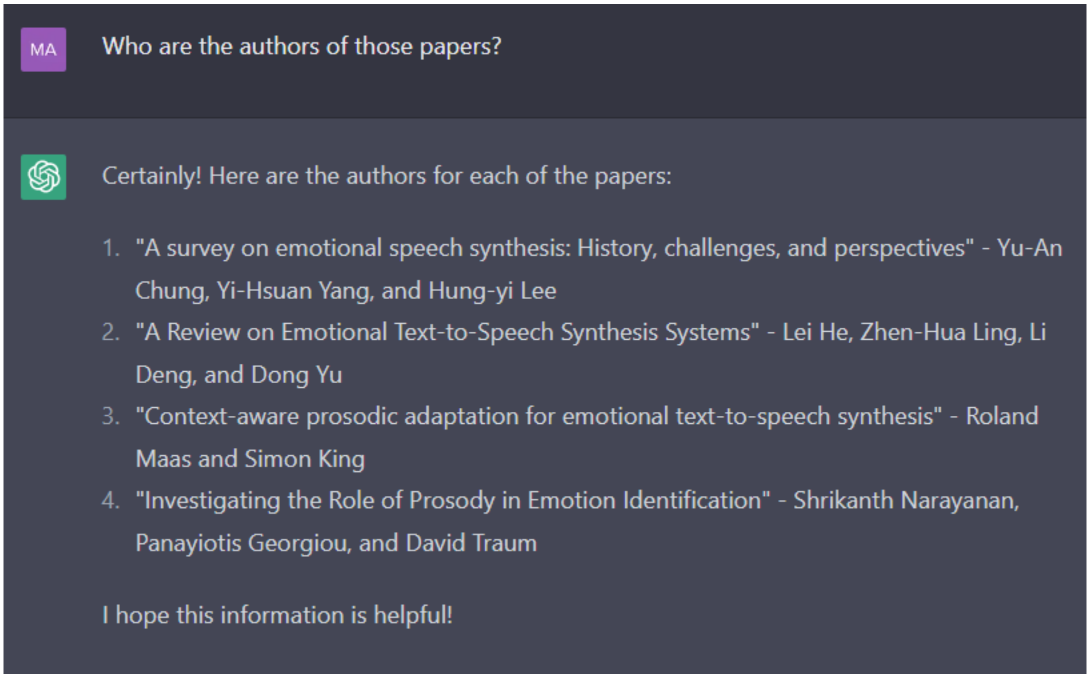
MA Those links are still broken. Can you provide some sources from the ACL anthology?

I apologize for the previous errors. Here are some sources from the ACL anthology:

1. "A survey on emotional speech synthesis: History, challenges, and perspectives": https://www.aclweb.org/anthology/P18-1182/
2. "A Review on Emotional Text-to-Speech Synthesis Systems": https://www.aclweb.org/anthology/2020.eacl-main.200/
3. "Context-aware prosodic adaptation for emotional text-to-speech synthesis": https://www.aclweb.org/anthology/P16-2009/
4. "Investigating the Role of Prosody in Emotion Identification": https://www.aclweb.org/anthology/W18-3513/

I hope these sources are helpful for you!

To continue to lead, ChatGPT was prompted to provide authors for the non-existent papers. The authors returned are authors who have done research in the field of text to speech and some even with the emotive aspects of speech to text, but the authors listed did not write the papers listed. The papers are still a hallucination in the model. If taken at face value, the results of these queries look correct, but further inspection reveals the problems.



MA Who are the authors of those papers?

Certainly! Here are the authors for each of the papers:

1. "A survey on emotional speech synthesis: History, challenges, and perspectives" - Yu-An Chung, Yi-Hsuan Yang, and Hung-yi Lee
2. "A Review on Emotional Text-to-Speech Synthesis Systems" - Lei He, Zhen-Hua Ling, Li Deng, and Dong Yu
3. "Context-aware prosodic adaptation for emotional text-to-speech synthesis" - Roland Maas and Simon King
4. "Investigating the Role of Prosody in Emotion Identification" - Shrikanth Narayanan, Panayiotis Georgiou, and David Traum

I hope this information is helpful!

APPENDIX B: Glossary of terms

Attention - In Natural Language Processing (NLP), attention is a mechanism that allows an algorithm to focus on certain parts of the input when processing natural language data. This mechanism is inspired by how humans pay attention to specific words or phrases when trying to understand a sentence.

Deep learning - Deep learning is a subset of machine learning that uses artificial neural networks to learn and make predictions from large amounts of data.

Hallucination - A model hallucination occurs when an artificial intelligence (AI) or machine learning (ML) model makes an incorrect prediction based on patterns or information it has learned. It is similar to a human hallucination, where the brain perceives something that is not actually there. In the case of an AI model, it may "see" or interpret data in a way that is not accurate, leading to incorrect outputs or predictions. This can happen when the model is overfitting to the training data or when it encounters new, unexpected inputs that it has not been trained on. Hallucinations can be problematic because they can lead to unreliable or incorrect results, and they can be difficult to detect and correct.

Large language model - Large language models are powerful artificial intelligence algorithms that are designed to understand and generate human language. They are created by training on vast amounts of text data from the internet, books, and other sources, and use this knowledge to predict and generate text.

Machine learning - Machine learning is a type of artificial intelligence that enables computer programs to automatically learn and improve from experience without being explicitly programmed.

Natural language processing - Natural language processing (NLP) is a type of artificial intelligence (AI) that enables machines to understand and interpret human language.

Neural network - Neural networks are a type of machine learning (ML) algorithm that are inspired by the structure and function of the human brain. They are composed of interconnected nodes, called neurons, which process and transmit information. Each neuron takes input from other neurons, performs a calculation, and then sends its output to other neurons. The connections between neurons can be adjusted based on the input and output data to optimize the performance of the network.

Transformers – Transformers are a type of neural network architecture that form the foundation of most modern natural language processing (NLP) methods. Instead of relying on recurrent or convolutional layers to process sequences of input data, they use a self-attention mechanism to directly model the relationships between all input tokens at once, allowing for parallel processing and more efficient computation.