# Pay attention: the high-speed evolution of NLP, and where it hits a wall

Fabio Cozman
Universidade de São Paulo (USP)
https://orcid.org/0000-0003-4077-4935
fgcozman@usp.br

Hugo Neri
Universidade de São Paulo (USP)
https://orcid.org/0000-0001-6065-4661
hugo.neri@hotmail.com

## ABSTRACT

This paper analyzes the evolution of attention-based models in Natural Language Processing (NLP) with an informal tone, starting from 2003 and culminating in the transformer architectures we know since 2017. We explain how transformers have managed to solve significant benchmarks for commonsense reasoning in Artificial Intelligence due to their pre-training. Further, we investigate the parallel between the concept of 'gist' in human language understanding, as proposed by Roger Schank, and the 'embeddings' now employed in machine learning. Towards the end of the paper, we discuss a well-known problem with these models, the so-called "hallucinations." This phenomenon highlights the models' struggle to discern fact from fiction, necessitating further research to mitigate its impact. We frame this issue in the context of David Lewis's work, arguing that it represents a fundamental challenge for language models.

**Keywords:** language models; deep learning; natural language processing; artificial intelligence.

# Preste atenção: a veloz evolução do processamento de linguagem natural e onde ela empaca

## RESUMO

Este artigo analisa a evolução dos modelos baseados em atenção no Processamento de Linguagem Natural (PLN) em um tom informal, começando em 2003 e culminando nas arquiteturas de "transformers" que conhecemos desde 2017. Explicamos como os "transformers" conseguiram resolver o importante "benchmark" para o raciocínio de senso comum em Inteligência Artificial devido ao seu pré-treinamento. Além disso, investigamos o paralelo entre o conceito de 'gist' ("o que realmente importa") na compreensão da linguagem humana, conforme proposto por Roger Schank, um veterano do PLN, e os "embeddings" agora empregados na aprendizagem de máquina. No final do artigo, discutimos um problema bem conhecido com esses modelos, as chamadas "alucinações". Este fenômeno destaca a luta dos modelos para discernir o fato de ficção, necessitando de mais pesquisas para mitigar seu impacto. Enquadramos essa questão no contexto do trabalho de David Lewis, argumentando que representa um desafio fundamental para os modelos de linguagem.

**Palavras-chave:** modelos de linguagem; aprendizado profundo; processamento de linguagem natural; inteligência artificial.

# 1. A NEW PREDATOR IN THE NLP LANDSCAPE

In the Darwinian landscape of machine learning, the advent of the attention mechanism led to the emergence of a new species of computational model for language processing. Similarly to the most adaptable organisms in the natural world, these models have evolved at first slowly, then dramatically, reshaping the landscape of natural language processing. The genesis of this novel species can be traced back to the concept of dynamically adjusting focus within sequences, mirroring how the human brain selectively attends to relevant *stimuli*. This idea appears in the groundbreaking paper "Neural Machine Translation by Jointly Learning to Align and Translate" by Bahdanau et al (2016).

Predating this, Bengio et al. (2003) laid the bedrock for the attention mechanism. They conceived a model that learned a distributed representation for each word and the probability function for word sequences simultaneously, employing a methodology evocative of how attention weights words based on their contextual relevance. Intuitively: they developed a method that gives each word a kind of 'tag' or 'score' that reflects its relationships with other words (the embedding of the word). At the same time, the method calculates how often certain words follow each other in a sequence. The resulting model does not truly understand the words or their meanings, but it can predict which words are likely to come next in a series (e.g., a text), perhaps the same way humans themselves predict language. Their exploration of vector space representations for words presaged the concept of word embeddings, a cornerstone in attention-based models. While their model did not fully implement the concept of dynamic input weighting based on the current context, the focus on a 'short list' of high-probability words hinted at an early, rudimentary form of attention.

In the evolution of NLP the advent of word embeddings has been a major milestone. You can think of word embeddings as a skill that these machine learning models acquired. It is similar to a tracking system, where each word gets a unique

numeric code, similar to animal tracks in the wild. Just like tracks can help predict the movements of an animal, these numeric codes help the models predict words that frequently appear together or are contextually related.

Imagine these numeric codes as having multiple dimensions, similarly to how animal tracks provide information about the size, species, or speed of the animal. In the case of word embeddings, each 'dimension' captures a different aspect of how a word interacts with other words in a language. The more dimensions we can track, the more information we have about the word's behavior in its natural habitat of language. This multi-faceted tracking system equips machine learning models with a powerful tool to navigate the complex world of natural language processing (Bengio et al., 2003; Mikolov et al., 2013a; Mikolov et al., 2013b; Pennington; Socher; Manning, 2014; Levy; Goldberg, 2014).

The attention mechanism, taking root from these foundational concepts, matured to surpass the constraints of previous NLP techniques. The mechanism can master long sequences, refine its understanding of context, and excel in parallel token processing. Equipped with such skills, the attention mechanism has tackled a diverse array of natural language problems that previous models struggled with.

A significant leap in the evolution of the attention mechanism came with the 2017 publication of Vaswani et al.'s paper, "Attention is All You Need" (2017). This paper began a radical shift from traditional architectures by positing that attention could form the bedrock of neural network architectures, thereby eliminating the need for recurrent and convolutional layers. This evolutionary leap instigated a dramatic reshaping, comparable to the emergence of a new apex predator that alters the balance of an ecological community. The Transformer model, a novelty introduced in that revolutionary paper, harnessed the power of the self-attention mechanism. This approach allowed the model to weigh and incorporate all tokens in the input sequence, irrespective of their positions, thus enabling it to establish dependencies between tokens far apart in

the sequence. The Transformer model also introduced a novel positional encoding scheme to account for word order, a crucial element in language understanding.

The Transformer model's scalability, coupled with its unparalleled capacity to model intricate patterns in data, paved the way for the development of larger and more powerful models. This led to a new generation of language models such as GPT (Radford et al., 2019) and Bert (Devlin et al., 2018), and Roberta (Liu et al., 2019) which have since achieved unprecedented performance across a broad spectrum of NLP tasks, from translation to text generation. Complex tasks such as machine translation, text summarization, and sentiment analysis succumbed to their prowess, and even longstanding challenges like the Winograd challenge, paraphrasing challenges, even arguments classification, have been surmounted by this apex predator.

## 2. THE 'PREY': NATURAL LANGUAGE CHALLENGES

The landscape of natural language processing is teeming with diverse 'prey,' or challenges, that continually test the prowess of the corresponding AI 'predators.' The Transformer model, the current apex predator, has shown remarkable adaptability and effectiveness, conquering various language challenges once thought insurmountable. For instance, the Transformer's expansive 'hunting ground' includes question answering, a task demanding a deep understanding of context and accurate information retrieval. On that terrain, the Transformer has demonstrated exceptional skill, exceeding previous performance standards on benchmarks such as the Squad dataset (Rajpurkar et al., 2016), a reading comprehension dataset consisting of questions posed by crowdworkers on a set of Wikipedia articles.

In the task of named entity recognition, earlier hunters such as Binary Decision Trees[1] (Bennett et al., 1997) and Hidden Markov Models[2] (Bikel et al., 1999) have been

[1] Binary Decision Trees are a form of machine learning model that makes decisions based on a series of yes/no questions about the data. Each node in the tree represents a condition or question about the data, and each branch represents the outcome of the question, leading to different paths down the tree. The final decision or prediction is made at the leaf nodes.

[2] Hidden Markov Models are statistical models where you observe a series of events but do not know the exact sequence of underlying states that led to those events. It uses probabilities to predict the sequence of hidden states.

superseded by Transformers. Even more recent ones like Long Short-Term Memory (LSTM[3]) (Lample et al., 2016), which despite being state-of-the-art for a time, have their relevance threatened. Transformers have left their mark by identifying and categorizing key information in text, such as the names of people, organizations, locations, and more. This task, analogous to tracking and classifying different types of prey, speaks to the Transformer's adeptness in understanding relationships between different words in a sentence. Transformers have also succeeded in the intricate territory of Natural Language Inference (NLI) with the Stanford Natural Language Inference (SNLI) corpus (Bowman et al., 2015). Much like navigating a complex terrain to capture elusive prey, NLI requires a delicate balance of language understanding and logical reasoning. Transformers, with their powerful mechanism for capturing long-range dependencies, have set new performance records on the Multinli dataset (Williams et al., 2018).

Transformers have also demonstrated their versatility by excelling in text classification, effectively categorizing text into predefined categories with remarkable accuracy.

Even the elusive Winograd family of problems, from WS-273 Davies (2016, 2017), Levesque (2011, 2013, 2017), and Levesque et al. (2012) to WinoGrande (Sakaguchi et al., 2019), possibly the wiliest of prey, has not remained unscathed by the Transformer. These problems of pronoun resolution, usually straightforward for humans but complex for machines, have been in essence solved by Transformers, thus illustrating their adeptness at nuanced and common-sense understanding.

As victors in the vast landscape of natural language processing, Transformers have vanquished many once-daunting challenges. They have shown remarkable dexterity in understanding human language, not by mimicking human cognitive processes but by displaying an uncanny knack for detecting patterns and relationships within colossal amounts of text data. This has resulted in pragmatic effects that

---

[3] The LSTM is a type of recurrent neural network that can remember or forget information over time.

outperform many traditional approaches and often approximate human-level performance.

Their victories are not confined to small battles; they have managed to master complex tasks like question answering, named entity recognition, and natural language inference. These achievements highlight the Transformer current dominance, but also pose new questions, akin to the mysteries that arise after every significant conquest.

The remaining natural language challenges are survivors indeed, but they are not merely remnants of a vanquished enemy. They represent new, exciting frontiers, challenges that expose the inherent complexity of human language and cognition, and the intricate nature of our world. These are not problems to be dismissed or understated, but intriguing puzzles that continue to captivate and challenge us. Despite their elusiveness, these problems do not pose insurmountable hurdles. They are akin to the hidden but not unreachable game in a savanna. Our engineering ingenuity, constantly evolving methodologies, and creative problem-solving can help us track these elusive targets.

The Transformers' victories, while astonishing, have also shaped the landscape of natural language processing in unexpected ways. They have allowed us to look into more abstract concepts and ideas, pushing the boundaries of what we once thought was achievable.

In the next sections of this paper, we will delve deeper into details behind the Transformers' success and explore the intriguing issue of hallucinations in language models. The hunt continues, but now it is about tracking down the elusive, complex beasts of the NLP savanna. These remaining challenges symbolize not just the limits of our current understanding, but also the fascinating possibilities for future breakthroughs.

## 3. TRANSFORMERS (PRE-) TRAINING IS COMMONSENSE TRAINING

Transformers have succeeded in cracking several challenging problems connected to commonsense reasoning. The journey towards understanding this success requires a foray into the roots of Bert and Roberta's training process, masked language modeling.

BERT, focusing on bidirectional language learning, masks a certain portion of words during its training phase. This adaptation mirrors a unique survival tactic in the linguistic wilderness — the cloze test. First introduced by Wilson Taylor in 1953, this test assesses communication effectiveness by asking individuals to fill in the blanks, completing the sentence much like closing gaps in a lineage. The parallel between Bert's masked language modeling and the cloze test is clear and striking.

Linguistic commonsense, the skill needed to complete a cloze unit, equips an individual with the ability to understand and complete mutilated sentences. This capability, which interacts with various enabling and obstructing elements, evolves into our comprehensive understanding of linguistic commonsense.

Given this, Bert's training can be viewed as a large-scale implementation of the cloze procedure, fostering the development of linguistic commonsense within a machine.

We should note that Winograd schemas (Davies, 2016; Davieset al., 2017) mimic cloze units in their environment. In fact, the Winograd Schema Challenge (WSC) demands this evolved skill, as it requires deducing anticipated words from their linguistic context. It is clear that Bert's evolutionary approach is well-suited for the challenges posed by the WSC.

Further into the linguistic wilderness, the Recognizing Textual Entailment (RTE) task, akin to the WSC, tests for Textual Entailment (TE) within sentences (Dagan, 2009). Despite RTE's reliance on inferring causal relationships, Levesque suggested the WSC as

a solution, introducing a mutation of double ambiguity, resulting in a cloze test-like environment (Levesque, 2011).

In this evolution-inspired narrative, the WinoGrande corpus and Roberta play key roles (LIU et al., 2019). Roberta, a more advanced version of BERT, has demonstrated significant leaps in performance within the WSC. This is attributed to superior pretraining involving larger corpora and more effective hyperparameters. In addition to the data sources used to train Bert, Roberta also incorporated the English portion of the CC-NEWS dataset, OpenWebText, and Stories from CommonCrawl.

The fine-tuning of Roberta using the Definite Pronoun Resolution Dataset (DPR) led to a leap in its fitness, reaching an impressive 83.1% accuracy in the WSC. This adaptation was further honed with the integration of WinoGrande, enhancing accuracy to an even more impressive 90.1%. Similar adaptive improvements were observed in the Pronoun Disambiguation Problems (PDP) dataset. Unexpectedly, the solution of the Choice of Plausible Alternatives (COPA) saw a substantial leap. Here, the expansive WinoGrande corpus provided a diverse environment conducive for training, with RoBERTa's performance soaring with the introduction of over 40k Winograd schemas.

While Roberta's progressive advancements and the vast WinoGrande environment are significant, some critics argue that these strides might be the result of evaluation anomalies and artifacts, rather than a genuine evolution in LLMs (Gururangan et al., 2018; Sakaguchi et al., 2019b). Regardless, the pivotal role of pretrained LLMs in enhancing word sense disambiguation is unquestionable, as demonstrated by the marked performance leap across different iterations of the WinoGrande project (Sakaguchi et al., 2019a).

Why are we confident that the masked language task during pretraining was the true driving force behind the impressive 90.1% performance observed in Sakaguchi et al. (2019b)? It is because an exceptionally large model such as GPT-3 did not achieve the anticipated stellar performance on the WSC: our tests resulted in a modest 67% accuracy. Performance improved for the even larger GPT-4, yielding an accuracy of 85.2%. While

commendable, this performance pales when compared to the models' sizes: GPT-3.5 has 175B parameters, and GPT-4 is even larger, though its exact size remains unknown (Openai, 2023). Conversely, Bert and Roberta have only 340 million parameters each. Indeed, size matters, but skill plays a far more crucial role. This represents a specialization within the current transformers landscape, with the Bert family rooted in pretraining that resembles a human reading task, while the GPT family simulates an oral task, where the next token prediction is paramount.

## 4. TRANSFORMERS CAPTURE, IN THEIR OWN WAY, THE GIST OF LANGUAGE

The notable cognitive scientist Roger Schank, in collaboration with Robert P. Abelson, substantially advanced the early field of Natural Language Processing (NLP) with their pioneering publication, "Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures" (1977). Their work established a foundational understanding of language processing and generation, both by humans and machines. Schank's distinctive view, centered on the intersection of AI and human cognition, brought a fresh perspective, even though it was not rooted in conventional machine learning or connectionist theories. Their perspective highlighted the necessity of cognitive frameworks in AI, prefiguring many subsequent discussions in NLP.

In the 1990s, Schank shifted his focus to education, thereby further augmenting his grasp of learning processes. The resonance and relevance of his initial work on AI persist in today's AI dialogue. His book from the '90s, "Tell Me a Story: Narrative and Intelligence," offers pertinent insights into our discourse on evolving more human-like AI. Schank postulates in that book that the essence of human intelligence and memory formation lies in storytelling and narratives, which suggests a unique approach to training AI systems to comprehend and create narratives, hence augmenting their 'human-like' attributes. This is especially pertinent to this paper as we examine how AI models deduce the 'gist' or underlying narrative patterns, akin to human cognition.

Schank proposes that storytelling forms the bedrock of human intelligence, asserting that narratives help us organize our memories, grasp our surroundings, and shape our life experiences. This extends beyond mere communication; narratives become the prism through which we interpret our world.

This perspective carries significant implications within Artificial Intelligence. It indicates that advancing towards a more human-like AI might depend on equipping machines with the capacity to understand and weave narratives. This is how large language models in fact function; they generate responses based on patterns and structures they have discerned from a massive spectrum of narratives in their training data.

Schank's theory emphasizes the concept of the "gist," the fundamental essence or core meaning of a story. According to Schank, we process and remember stories by distilling these main themes or schemas, abstract structures representing our knowledge and expectations. When we listen to a story, we do not remember every detail but retain the gist — the basic structure and the key elements that lend meaning to the story. To an extent, AI language models replicate this process.

This idea of a "gist" in Schank's theory exhibits parallels to "embeddings" in machine learning. Embeddings are mathematical constructs where similar entities are grouped closely. In the context of natural language processing, words or phrases with similar meanings are represented as vectors in a multidimensional space, situated near each other. Just as the "gist" encapsulates a story's essential meaning, these embeddings distill the fundamental connotation of a word or phrase.

The ability of embeddings to capture semantic similarities is crucial in recognizing synonyms across different dialects or jargons. For example, a sophisticated language model can comprehend that 'car' and 'automobile' or 'solicitor' and 'lawyer' carry similar meanings, even when used in varying contexts. This ability is critical in applications like information retrieval or machine translation, as it allows accurate mapping of words and phrases across languages.

Nonetheless, embeddings, like all tools, have their constraints. Their efficacy hinges on the quality and diversity of the training data. If certain words, phrases, or concepts are underrepresented, the embeddings might not accurately encapsulate their meanings or subtleties. This underlines the significance of a comprehensive, diverse, and high-quality dataset for training more nuanced and effective language models.

## 5. LLMs CAN EXPLORE THE UNREAL

Large language models such as GPT-4 have the ability to engage with hypothetical scenarios and counterfactuals, offering a fascinating glimpse into the world of "what ifs". This ability, integral to human cognition, allows for exploration of uncharted territory, evaluation of different possibilities, and anticipation of future scenarios. Advanced language models like GPT-4 can create responses to these hypothetical situations based on patterns learned from vast text data. However, their responses lack the genuine emotional understanding or empathy that underlies human responses.

The concept of Counterpart Theory by David Lewis provides a useful perspective here. In that theory, entities in different possible worlds can be 'counterparts' with similar characteristics but not identical (LEWIS, 1968). The responses generated by AI models to hypothetical scenarios can be considered 'counterparts' to human responses — they imitate human behavior but remain fundamentally different due to the absence of subjective experiences, emotions, and ethical comprehension.

Lewis's notion of truth within fictional narratives is another significant aspect. He suggests that fictional stories create their own 'possible worlds', and statements about these worlds can be considered true within the confines of those worlds (Lewis, 1978). Language models, using the patterns, relationships, and rules they have learned, can generate narratives that are coherent within the possible world of a given fiction.

The Principle of Minimal Departure proposed by Lewis states that when a story leaves certain details unspecified, we fill these gaps based on our real world, assuming they do not contradict the narrative (Lewis, 1968). Language models often adopt a

similar approach when generating text, filling in the blanks with patterns and facts learned from their training data. The fact that truth in fiction is context-dependent, according to Lewis, presents a significant challenge and opportunity for large language models. While they are capable of producing varied responses depending on context, their sensitivity to context relies more on pattern recognition than on genuine understanding.

Future developments in this area may answer intriguing questions like: Can large language models generate more sophisticated responses to counterfactuals? Will they better leverage context to navigate the nuances of truth within fiction? As they continue to evolve, will they further refine their mimicry of human-like understanding, or will they develop a unique way of interpreting these issues?

## 6. PARALLEL UNIVERSES

David Lewis's philosophy is based on the concept of possible worlds, which represent comprehensive ways things could have been. This concept resonates with the universe of AI, especially language models like GPT-4, which live in a world of linguistic data devoid of physical matter. Their 'possible world' is constructed from patterns, statistical associations, and symbolic relationships encoded in the training data.

Lewis's philosophy espouses modal realism about possible worlds, considering each of these worlds as real as our actual world. This parallels with the 'world' of a language model which is constituted by linguistic data and relationships (Lewis, 1986). From Lewis's perspective, a plethora of possible worlds exist, and our actual world is merely one of these realities. This is similar to the functioning of language models like GPT-4, wherein each unique combination of input parameters and model states can lead to the generation of a different 'reality' or output.

Lewis further posits that these possible worlds exist independently and do not causally interact. Despite this isolation, we can conceptualize and communicate about these alternate worlds, enabling the creation of fiction. AI language models function in a

similar manner, treating each input-output pair separately, with no causal interaction between different pairs. However, they can produce a variety of contextually appropriate responses, crafting narratives much like a fictional story. In the realm of AI, Lewis's counterpart theory can be related to the fact that each unique output from a model for a given input can be viewed as 'counterparts' — not identical but similar representations.

## 7. HALLUCINATIONS IN LANGUAGE MODELS: A PERENNIAL RISK? TRANSFORMERS CANNOT DISTINGUISH BETWEEN TRUTH AND FICTION, ACTUAL AND POSSIBLE WORLDS

Hallucination in the context of LLMs such as GPT-4 refers to the generation of facts that are not related to a given input nor directly derived from the training data. This phenomenon typically surfaces when the model is tasked with filling in gaps or extrapolating on ambiguous inputs. As it strives to generate human-like text, it draws on its extensive training to make educated guesses, sometimes leading to outputs that, while seemingly plausible, are not necessarily grounded in reality.

A pivotal issue in this process lies in the LLM's inability to discern fiction from fact. These models are trained on a vast corpus of text data that comprises a blend of factual and fictional content. Absent an inherent understanding of the difference, they may inadvertently 'hallucinate' elements from their training data, weaving fictional elements into contexts that call for strictly factual information, and vice versa.

For instance, when asked about a historical event, an LLM might generate a response that blends factual data with fictional elements derived from a novel set in the same era. Though the generated text may appear coherent and contextually appropriate, it may misrepresent the actual event, thus generating a 'hallucination.'

The potential risks associated with such hallucinations are significant. In critical applications such as medical advice or legal counsel, hallucinations may lead to erroneous decisions with grave consequences. For example, an LLM trained on a mixture of real medical textbooks and science fiction novels might 'hallucinate' while providing medical advice, yielding recommendations that are scientifically unfounded and

potentially harmful. In the same vein, a LLM can be tricked to do not distinguish a fictional conversation to a real one, allowing it to produce harmful speech.

So, how do we mitigate these hallucinations? Strategies include refining the quality and diversity of training data, implementing fact-checking mechanisms during the output generation process, and exploring techniques to limit the LLM's propensity to generate text beyond its training data. However, completely eliminating hallucinations remains a formidable (if not impossible) challenge, primarily due to the inherent design and training methods of these models.

Hallucination, while a challenge, could also be perceived as a perennial trait of language models like GPT-4, leading to an intriguing paradox. Their strength — generating human-like text — can also be their Achilles' heel when the boundaries blur between fiction and reality. As we continue to navigate this complex landscape, these 'hallucinations' stand not just as a risk to be managed, but also as a compelling puzzle driving us towards the development of more sophisticated models. Models that can better distinguish fiction from reality, effectively reducing hallucinations, all the while refining our understanding of the vast and nuanced terrain of artificial intelligence.

## 8. FINAL REMARKS

Large Language Models (LLMs) have reached unprecedented heights in language-related tasks, often exhibiting capabilities that rival, or even surpass, human proficiency. The breadth of their linguistic knowledge and their ability to adapt to various contexts allow them to not only comprehend and generate language but to do so in a way that encapsulates the essence, or the "gist," of the discourse at hand.

LLMs leverage 'embeddings', a powerful representation of words or phrases as vectors in a high-dimensional space. These embeddings are trained to capture the semantic essence of words, facilitating translation between languages by mapping similar concepts together. However, LLMs push this capability even further by translating not just across different languages, but also between audiences. This means

they can take a concept expressed in specialized jargon and translate it into plain language, making complex topics accessible to a wider audience.

However, the prowess of LLMs is rooted solely in the realm of language. They exist in a world of linguistic data, devoid of physical matter. This inherent characteristic shapes their limitations as well as their strengths. A significant limitation that comes to the fore is the phenomenon of 'hallucination.' In the context of machine learning, hallucination refers to the generation of details that are not present in the input or training data. When an LLM faces a linguistic situation with missing or ambiguous information, it often 'fills in the gaps,' which may lead to outputs that are unrelated to the original context or not based in reality. This makes the task of mitigating such undesired outputs an external endeavor, not inherent to the model itself.

The enormous size of LLMs, while contributing to their extraordinary capabilities, also poses challenges. These models require vast computational resources for training and operation, making them difficult to manage and potentially raising issues related to efficiency and environmental sustainability. Moreover, the complexity of these models can make understanding and predicting their behavior increasingly challenging. However, the discussion around these concerns is expansive and deserves a dedicated exploration, potentially opening new avenues of research and innovation.

## REFERENCES

BAHDANAU, D.; CHO, K.; BENGIO, Y. **Neural Machine Translation by Jointly Learning to Align and Translate. 2016**. Disponível em: https://arxiv.org/abs/1409.0473.

BENGIO, Y. et al. A neural probabilistic language model. **Journal of Machine Learning Research**, v. 3, n. Feb, p. 1137-1155, 2003.

BENGIO, Y.; DUCHARME, R.; VINCENT, P.; JANVIN, C. **A neural probabilistic language model. Journal of Machine Learning Research**, v. 3, p. 1137-1155, 2003.

BENNETT, S.W.; AONE, C.; LOVELL, C. Learning to tag multilingual texts through observation. In: SECOND CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2., 1997, Providence. Proceedings. Providence: Morgan Kaufmann Publishers, Inc, 1997. p. 109-116.

BOWMAN, S. R.; ANGELI, G.; POTTS, C.; MANNING, C. D. A large annotated corpus for learning natural language inference. 2015. Available at: http://arxiv.org/abs/1508.05326.

BROWN, T. B.; et al. **Language models are few-shot learners**. ArXiv:2005.14165, 2020.

DAGAN, Ido. **Recognizing textual entailment**: Rational, evaluation and approaches. Natural Language Engineering, v. 15, n. 4, p. i-xvii, 2009.

DAVIES, Ernest. **Winograd schemas and machine translation**. 2016. Disponível em: https://arxiv.org/abs/1608.01884.

DAVIES, Ernest; MORGENSTERN, Leora; ORTIZ, Charles L. The first Winograd Schema Challenge at IJCAI-16. AI Magazine, v. 38, n. 3, p. 97-98, 2017.

DEVLIN, J.; CHANG, M. W.; LEE, K.; TOUTANOVA, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018. Available at: https://arxiv.org/abs/1810.04805.

GURURANGAN, Suchin; et. al. Annotation Artifacts in Natural Language Inference Data. 2018. Available at: https://arxiv.org/abs/1803.02324.

LEVESQUE, Hector. The Winograd Schema Challenge. In: AAAI SPRING SYMPOSIUM, 2011, Palo Alto. Anais. Palo Alto: AAAI, 2011.

LEVESQUE, Hector. On our best behaviour. In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, 2013, Beijing. Anais. Beijing: IJCAI, 2013.

LEVESQUE, Hector. Common Sense, the Turing Test, and the Quest for Real AI. Cambridge, Massachusetts: The MIT Press, 2017.

LEVESQUE, Hector; DAVIES, Ernest; MORGENSTERN, Leora. The Winograd Schema Challenge. In: PRINCIPLES OF KNOWLEDGE REPRESENTATION AND REASONING, 2012, Rome. Proceedings. Rome: KR, 2012.

LEVY, O.; GOLDBERG, Y. Neural word embedding as implicit matrix factorization. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS. Anais. 2014. p. 2177-2185.

LEWIS, D. **Counterpart theory and quantified modal logic. Journal of Philosophy**, v. 65, n. 5, p. 113-126, 1968.

LEWIS, D. On the Plurality of Worlds. Oxford: Blackwell, 1986.

LEWIS, D. Truth in Fiction. American Philosophical Quarterly, v. 15, n. 1, p. 37-46, 1978.

LIU, Y. et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. 2019. Available at: https://arxiv.org/abs/1907.11692.

MIKOLOV, T. et al. Distributed representations of words and phrases and their compositionality. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, Anais. 2013. p. 3111-3119.

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. 2013. Disponível em: https://arxiv.org/abs/1301.3781.

OPEN AI. GPT4 Technical Report. ArXiv:2303.08774, 2023.

PENNINGTON, J.; SOCHER, R.; MANNING, C. Glove: Global vectors for word representation. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2014. Proceedings. 2014. p. 1532-1543.

RADFORD, A.; WU, J.; CHILD, R.; LUAN, D.; AMODEI, D.; SUTSKEVER, I. Language models are unsupervised multitask learners. OpenAI Blog, v. 1, n. 8, 2019.

RADFORD, A.; WU, J.; CHILD, R.; LUAN, D.; AMODEI, D.; SUTSKEVER, I. Language Models are Unsupervised Multitask Learners. OpenAI Blog, v. 1, n. 8, 2019. Available at: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

RAJPURKAR, Pranav; ZHANG, Jian; LOPYREV, Konstantin; LIANG, Percy. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2016, Austin, Texas. Anais. Austin, Texas: Association for Computational Linguistics, 2016. p. 2383-2392.

RYAN, M. L. **Possible worlds, artificial intelligence, and narrative theory**. Bloomington: Indiana University Press, 1991.

SAKAGUCHI, K.; LE BRAS, R.; BHAGAVATULA, C.; CHOI, Y. Winogrande: An adversarial Winograd Schema Challenge at scale. In: AAAI-20 TECHNICAL TRACKS 5, 34, 2019. p. 05.

SCHANK, R.; ABELSON, R.P. **Scripts, plans, goals and understanding**: An inquiry into human knowledge structures. New Jersey: Erlbaum, 1977.

SCHANK, R.C. **Tell Me a Story**: A New Look at Real and Artificial Memory. 1st ed. New York: Atheneum, 1990.

TAYLOR, W. "Cloze procedure": A new tool for measuring readability. Journalism Quarterly, v. 30, n. 4, p. 415-433, fall, 1953.

VASWANI, A. et al. **Attention is all you need**. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS. Anais, 2017. p. 5998-6008.

WILLIAMS, A.; NANGIA, N.; BOWMAN, S. **A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference.** In: CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES, 2018, New Orleans. Anais. New Orleans: Association for Computational Linguistics, 2018. p. 1112-1122.