UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE COMUNICAÇÃO E EXPRESSÃO
DEPARTAMENTO DE LITERATURA E LÍNGUAS VERNÁCULAS
CURSO LETRAS - LÍNGUA PORTUGUESA E LITERATURAS

Laiara Machado Serafim

Do Humano aos LLMs: Apontamentos Sobre o Processamento de Linguagem

Florianópolis
2025

Laiara Machado Serafim

Do Humano aos LLMs: Reflexões Sobre o Processamento de Linguagem

Trabalho de Conclusão de Curso submetido ao curso de Letras - Língua Portuguesa e Literatura do Centro de Comunicação e Expressão da Universidade Federal de Santa Catarina como requisito parcial para a obtenção do título de Bacharel(a) em Letras – Língua Portuguesa e Literaturas.

Orientador(a): Prof.(a), Dr.(a) Roberta Pires de Oliveira

Florianópolis
2025

Serafim, Laiara Machado
Do Humano aos LLMs : : Apontamentos Sobre o Processamento de Linguagem /Laiara Machado Serafim ; orientadora, Roberta Pires de Oliveira, 2025.
38 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro de Comunicação e Expressão, Graduação em Letras - Língua Portuguesa, Florianópolis, 2025.

Inclui referências.

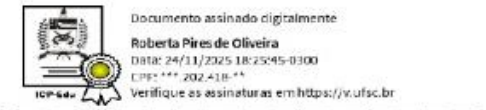
1. Letras - Língua Portuguesa. 2. Processamento de Linguagem. 3. Large Language Models. 4. Línguas Naturais. I. Pires de Oliveira, Roberta. II. Universidade Federal de Santa Catarina. Graduação em Letras - Língua Portuguesa. III. Título.

Laiara Machado Serafim

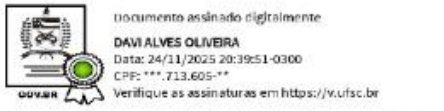
Do Humano aos LLMs: Apontamentos Sobre o Processamento de Linguagem

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do título de Bacharel e aprovado em sua forma final pelo Curso de Letras.

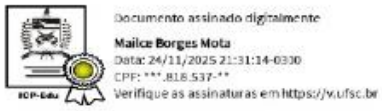
Florianópolis, 24 de novembro de 2025.



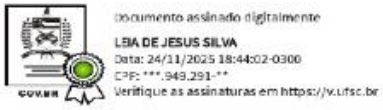
ROBERTA PIRES DE OLIVEIRA
Presidente e orientador(a)



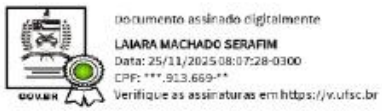
DAVI ALVES OLIVEIRA
Membro titular



MAILCE BORGES MOTA
Membro titular



LEIA DE JESUS
Suplente



LAIARA MACHADO SERAFIM
Acadêmica

AGRADECIMENTOS

À minha mãe, a mulher mais forte e corajosa que eu já conheci.

Ao meu pai, sempre haverá uma parte de você em mim.

Às minhas irmãs, que geraram os maiores amores do mundo – Osório, Micaela, Gregório e Melissa.

Ao Arthur, que sempre acredita em mim quando eu mesma não consigo. E está aqui, sempre.

À Luci e ao Tita, que me deram um terceiro lar durante esse percurso.

À Roberta Pires de Oliveira, minha orientadora, exemplo de pesquisadora e professora. Este trabalho não seria nada sem você.

Aos professores que, desde a minha infância, me fizeram acreditar que a educação muda vidas. Não tenho certeza sobre todas, mas tem mudado a minha.

“Palavras, na minha não tão humilde opinião, são nossa inesgotável fonte de magia.”
– Harry Potter

RESUMO

Este trabalho realiza uma primeira reflexão comparativa entre o processamento de linguagem humano e os *Large Language Models* (LLMs), com base em uma revisão bibliográfica que integra contribuições da linguística, neurociência, filosofia da mente e ciência da computação. Partindo de marcos teóricos como a Gramática Universal de Chomsky (1955) e a pobreza do estímulo, o estudo contrasta a hipótese da faculdade linguística inata humana – recursiva, criativa e ancorada em experiências sensoriais e contextuais – com a arquitetura probabilística e baseada em dados dos LLMs, característica do cognitivismo. A hipótese central é que, embora os LLMs tenham alcançado notável capacidade de geração de texto (superando, em alguns contextos, o Teste de Turing (1950)), sua operação permanece fundamentalmente estatística, destituída de compreensão semântica, intencionalidade ou criatividade genuína. Por meio da análise de experimentos recentes (Linzen & Leonard, 2018; Amouyal *et al*., 2024; Houghton e*t al*., 2023), demonstra-se que os modelos computacionais falham em replicar mecanismos cognitivos humanos, como a generalização a partir de dados limitados, a interpretação contextualizada e a ancoragem referencial no mundo real. Conclui-se que a diferença essencial reside na natureza experiencial da cognição humana, que integra linguagem, corpo e mundo, em contraste com o funcionalismo computacional dos LLMs, que opera por recombinação estatística de padrões linguísticos. A pesquisa sugere, portanto, que a capacidade humana de criar significado novo, imprevisível e intencional constitui uma fronteira ainda intransponível para a inteligência artificial.

Palavras-chave: Processamento de Linguagem; *Large Language Models*; Línguas Naturais.

ABSTRACT

This work undertakes a preliminary comparative reflection on human language processing and Large Language Models (LLMs), based on a literature review that integrates contributions from linguistics, neuroscience, philosophy of mind, and computer science. Departing from theoretical frameworks such as Chomsky's Universal Grammar (1955), and the poverty of the stimulus, the study contrasts the hypothesis of an innate human linguistic faculty—recursive, creative, and anchored in sensory and contextual experiences—with the probabilistic, data-driven architecture of LLMs, characteristic of cognitivism. The central hypothesis is that, although LLMs have achieved remarkable text-generation capabilities (surpassing, in some contexts, the Turing Test (1950)), their operation remains fundamentally statistical, devoid of semantic understanding, intentionality, or genuine creativity. Through the analysis of recent experiments (Linzen & Leonard, 2018; Amouyal *et al*., 2024; Houghton *et al*., 2023), it is demonstrated that computational models fail to replicate human cognitive mechanisms, such as generalization from limited data, contextualized interpretation, and referential anchoring in the real world. It is concluded that the essential difference lies in the experiential nature of human cognition, which integrates language, body, and world, in contrast with the computational functionalism of LLMs, which operates through the statistical recombination of linguistic patterns. The research suggests, therefore, that the human capacity to create new, unpredictable, and intentional meaning constitutes a frontier still insurmountable for artificial intelligence.

Keywords: Natural Language Processing; Large Language Models; Natural Languages.

SUMÁRIO

LISTA DE FIGURAS

INTRODUÇÃO

A linguagem é considerada a capacidade mais distintiva da espécie humana, servindo como uma das principais chaves para a nossa evolução e desenvolvimento cultural (Pinker, 1994). Por meio dela, transmitimos experiências, preservamos memórias e projetamos possibilidades para o futuro. Diferente de outros sistemas de comunicação encontrados na natureza, a linguagem humana possui uma complexidade estrutural e simbólica que possibilita abstração, criatividade e a elaboração de conceitos (Schlenker *et al*, 2016). É justamente essa capacidade que nos torna aptos para criar e organizar, compartilhar significados e transformar a realidade em que vivemos. Isso significa que podemos combinar um número finito de sons para criar um número infinito de sentenças, nos permitindo expressar ideias novas, complexas e abstratas sobre o passado, o presente e o futuro, o possível e o impossível.

A linguagem é tão central à nossa essência que, ao buscar replicar a inteligência humana, nosso foco não se volta primariamente a aspectos de forma, mas à capacidade linguística que permite a comunicação. Desde o teste de Alan Turing (1950), buscou-se criar máquinas que pudessem replicar a inteligência humana e, para “passar no teste”, era necessário conseguir dialogar de forma indistinguível da humana. Ou seja, não buscamos apenas replicar a nossa aparência, mas sim a nossa capacidade mais fundamental: a de usar a linguagem para raciocinar, criar e conectar. O desenvolvimento de modelos como os chamados Grande Modelo de Linguagem – do inglês *Large Language Model* (LLM) – (como o Chat GPT, Bert e Gemini), que servirão como a base contrastiva entre o processamento linguístico[1] em humanos neste trabalho, demonstra o quão longe chegamos nessa investigação.

Ao longo das últimas décadas, tanto a linguística quanto a ciência da computação – esta iniciada formalmente na década de 60 – têm se debruçado sobre o desafio de compreender e modelar os mecanismos envolvidos no processamento da linguagem. Se, por um lado, o cérebro/mente humano realiza esse processo de forma notavelmente eficiente desde os primeiros anos de vida, por outro, as máquinas vêm, cada vez mais, gerando linguagem de forma surpreendente e indistinta do

[1] Elemento que será definido posteriormente ainda nessa seção.

humano (Mei *et al*, 2024) por meio de sistemas computacionais sofisticados. A emergência e o rápido desenvolvimento da Inteligência Artificial (IA), especialmente no campo da linguagem, com os LLMs que surgem por volta de 2017, estabeleceram uma interface direta com a linguística e a neurociência. Esse breve trabalho de apontamento/reflexão, em conjunto com diferentes áreas do saber, que constituem a chamada Ciências Cognitivas em sentido amplo, levanta questões como “como o cérebro/mente humana e os sistemas de LLMs se comportam no processo de interpretação/decodificação e geração de linguagem?”, “quais atividades são realizadas nesse processo?”, e, mais crucialmente, se há e quais as razões para as distinções observáveis entre eles.

Embora ambos os sistemas demonstrem notável capacidade de lidar com a linguagem, como proposto pelo teste Turing que será abordado em seção posterior, proposto ainda em 1950, o qual mostrou que as máquinas são capazes de “imitar” uma conversa humana (Mei *et al*, 2024), suas arquiteturas subjacentes e princípios de aprendizagem parecem divergir significativamente. E é aqui que se encontra o cerne deste trabalho: uma reflexão, ainda que breve, com base na literatura apresentada, sobre as características de cada sistema que nos revelam que, por mais que Turing e seu teste tenham sido bem sucedidos na ideia de criar uma máquina capaz de replicar, em partes, a fala humana, há diferenças fundamentais em sua capacidade e em sua natureza.

Na base dessa reflexão está a famosa dicotomia entre cérebro, a substância da matéria, e mente, a substância do espírito, no sentido cartesiano, ou, como prefiro abordar, que se baseia na experiência. Do ponto de vista teórico da Linguística Gerativa, central à presente análise, a distinção entre "cérebro" e "mente" é tratada por seu fundador Noam Chomsky como um mistério, contornado pela noção de "Faculdade da Linguagem" enquanto um sistema computacional biológico. Essa abordagem não se orienta pela investigação direta do substrato neuroanatômico (o cérebro como órgão material), nem por uma concepção dualista da mente como entidade imaterial. Em vez disso, postula-se a existência de uma estrutura mental – a Gramática Universal (GU) (Chomsky, 1965), um princípio computacional fundamental que é recursivo e inato, o qual constitui a arquitetura inicial e geneticamente determinada do sistema linguístico. Essa estrutura é um componente da mente no sentido representacional, mas é, também, um objeto biológico, instanciado na estrutura do cérebro. Dessa forma, este trabalho não tratará tão somente da mente

introspectiva ou do cérebro em sua fisiologia, mas sim da natureza formal e das propriedades abstratas deste componente específico da cognição humana, inferidas a partir dos seus produtos linguísticos.

Ainda, para os fins deste trabalho, o termo "processamento de linguagem" será delimitado como a capacidade cognitiva e/ou computacional de gerar e interpretar/decodificar proferimentos linguísticos em conformidade com o sistema de regras de uma língua natural. Isso envolve, de um lado, o processo produtivo de selecionar e combinar elementos lexicais mínimos para a produção de sentenças gramaticalmente bem-formadas e semanticamente coerentes com a situação de proferimento. De outro lado, inclui o processo receptivo de decodificar tais sentenças, atribuindo-lhes uma representação sintática da qual se deriva composicionalmente seu significado, com base no conhecimento linguístico e no contexto (Dell, 1986). Este duplo movimento – de interpretação e geração – constitui o núcleo do objeto aqui investigado, servindo como eixo comparativo central entre a Faculdade da Linguagem humana e as operações realizadas pelos sistemas de LLMs. Esta concepção de processamento alinha-se com os objetos de estudo da Psicolinguística (Traxler & Gernsbacher, 2011; Maia, 2015), que investiga, inclusive, os mecanismos de produção e compreensão em tempo real (*online*), e pressupõe a existência de um sistema algorítmico de conhecimento linguístico – uma gramática internalizada – que torna tais processos possíveis. Ao adotar essa perspectiva, que entende a linguagem como um sistema computacional (Pinker, 1994), estabelece-se um eixo teórico comum que permite contrastar a natureza biológica e cognitiva do processamento humano com a arquitetura probabilística dos LLMs.

Nesse cenário, vemos surgir mais recentemente diferentes experimentos empíricos que buscam verificar se as máquinas, tal como estão concebidas hoje em dia, comportam-se ou não de maneira análoga aos seres humanos em diferentes tarefas linguísticas. Não se trata agora, somente, de “se passar por um falante”, mas de gerar e operar fenômenos linguísticos inerentes à fala humana, como ambiguidade, pressuposição, ironia etc. Trata-se de uma capacidade que, conforme os experimentos que serão aqui analisados, não parece possível de replicar ou serem apreendidas de forma precisa apenas com padrões estatísticos, isto é, apenas indutivamente. A comparação entre esses dois sistemas e a reflexão sobre os resultados obtidos nos testes já existentes na área ((Linzen & Leonard (2018); Moraes *et a*l (2024); Amouyal, Meltzer-Asscher & Berant (2024)), buscará lançar luz sobre as

semelhanças e as diferenças estruturais, funcionais e conceituais entre os modos como o humano e a máquina lidam com a linguagem.

A REVOLUÇÃO COGNITIVA

A linguagem é considerada um dos principais indícios da singularidade humana. Para René Descartes (1637), nenhum animal ou máquina seria capaz de utilizar a linguagem de forma criativa e ilimitada, como fazem os seres humanos. Essa capacidade de gerar expressões novas e adequadas a contextos variados revelaria a existência de uma substância pensante — a *res cogitans* — distinta da dimensão puramente mecânica do corpo. A linguagem, portanto, surgia não apenas como meio de comunicação, mas como prova do pensamento e da razão.

Mas foi Wilhelm Humboldt (1836, tradução de 2009) um dos primeiros pensadores a propor uma relação profunda entre linguagem e cognição. Sua concepção da linguagem como atividade formadora do pensamento antecipou discussões contemporâneas sobre a relação entre estruturas linguísticas e processos cognitivos. Sua noção de que cada língua expressa uma visão de mundo distinta serviu como ponto de partida para pensar o modo como o cérebro/mente humano — e, por contraste, os modelos computacionais — organizam e processam a linguagem. Segundo Corrêa (2006):

> “Além de Humboldt conceber um sistema gerativo universal que daria conta da produtividade das línguas, viria apresentar um tipo de solução para a questão da aquisição do conhecimento linguístico ante a variabilidade das línguas. Humboldt assume uma disposição natural para a língua no ser humano, assim como lhe atribui uma capacidade para aquisição de qualquer língua. A variabilidade das línguas estaria nos “meios” (o que pode ser entendido como sua expressão fonológica e morfológica) e assume “limites” ou restrições para sua realização.” (Corrêa, 2006, p. 8)

Segundo a autora, um sistema de geração de linguagem sem restrições internas teria uma carga computacional insustentável. Dessa forma, o cérebro humano precisaria de um "espaço de hipóteses" delimitado para processar e adquirir a língua de forma eficiente. As restrições inatas (por exemplo, a estrutura hierárquica da sentença, a capacidade de atribuir papéis semânticos e, como viríamos a saber, a recursividade) canalizam o processo de aquisição, guiando a criança a buscar padrões específicos nos dados linguísticos aos quais ela é exposta. Na produção da

fala, o falante não combina palavras aleatoriamente. As "restrições" humboldtianas seriam a gramática interna que, de forma rápida e inconsciente, limita as combinações possíveis, gerando apenas sequências bem-formadas e significativas na sua língua. É a diferença entre gerar "O gato persegue o rato" e a agramatical *Persegue rato gato o, em português, mas talvez possível como estrutura em outras línguas. De fato, a criatividade e capacidade de criar sentenças completamente novas é plenamente possível, mas opera dentro de um quadro predefinido.

No campo da psicologia, o behaviorismo clássico de Frederic Skinner exerceu forte influência sobre os estudos da linguagem. Em sua obra *Verbal Behavior* (1957), Skinner concebia a linguagem como um comportamento aprendido por meio de mecanismos de condicionamento e reforço. Para ele, as palavras e sentenças não passariam de respostas a estímulos ambientais, moldadas pela repetição e pela associação com consequências positivas ou negativas. Assim, o foco recaía na observação objetiva do comportamento linguístico, deixando de lado as estruturas mentais ou processos internos. Essa perspectiva, embora coerente com a proposta behaviorista de eliminar explicações mentalistas, mostrava-se limitada para dar conta da complexidade e da criatividade inerentes ao uso da linguagem humana. Além disso, até meados do século XX, a linguística norte-americana era fortemente influenciada por Leonard Bloomfield (1933), cuja proposta privilegiava a descrição formal e empírica das línguas, com ênfase nos aspectos observáveis de sua estrutura e na rejeição de hipóteses de caráter mentalista.

Ao longo do século XX, a concepção de linguagem passou por transformações significativas, especialmente com o advento da chamada Revolução Cognitiva. Nesse contexto, surgiram os primeiros estudos da psicolinguística, ainda nos anos 1950, voltados para investigar experimentalmente a produção e a compreensão da linguagem.

Com a emergência do gerativismo linguístico, proposto por Noam Chomsky, a compreensão da linguagem passou a ser reinterpretada. Em *Syntactic Structures* (1957) e em sua crítica a *Verbal Behavior* (1959), Chomsky estruturou a ideia de que os seres humanos possuem uma faculdade inata da linguagem, capaz de gerar, a partir de um conjunto finito de regras, uma quantidade infinita de sentenças. Esse enfoque inovador transformou a psicolinguística, que passou a se consolidar como o estudo dos processos mentais internos que permitem a aquisição, o uso e a compreensão da linguagem. Essa perspectiva inaugurou uma nova maneira de

pensar a linguagem: não mais como um reflexo do ambiente, mas como uma manifestação da capacidade criativa da mente humana.

Décadas depois, nos anos 1980, surge a linguística cognitiva, em parte como reação ao gerativismo, com a publicação de “Metaphors we live by” de George Lakoff e Mark Johnson. Essa corrente entende a linguagem como um fenômeno emergente da cognição geral, da experiência e da interação social, enfatizando o papel do contexto, da percepção e das habilidades cognitivas na construção de significado. Diferentemente do gerativismo, que postula uma faculdade inata da linguagem, a linguística cognitiva busca compreender a linguagem como um produto de processos cognitivos interligados.

A partir dessas linhas de pesquisa, o interesse deste trabalho volta-se à estrutura da mente e aos processos que possibilitam a aquisição, o uso e a compreensão da linguagem, dialogando, assim, mais intensamente com áreas como a psicolinguística e a filosofia da mente. Torna-se, cada vez mais, importante ampliar a investigação sobre como o cérebro/mente humano processa a linguagem, constrói significados e se adapta a contextos comunicativos complexos.

CHOMSKY E O GERATIVISMO LINGUÍSTICO

É nesse contexto, em contraposição às concepções behavioristas predominantes na época, que postulavam a linguagem como um conjunto de respostas condicionadas a estímulos externos, que Chomsky propôs uma abordagem radicalmente diferente para aquele momento, centrada nas capacidades intrínsecas da mente humana. Nesse cenário, ele defendeu que a linguagem não poderia ser explicada unicamente como um sistema de hábitos ou como uma organização de formas externas, mas que deveria ser compreendida como manifestação de uma competência mental subjacente, por causa da criatividade humana que não pode ser explicada por mecanismos de estímulo-resposta. Essa reformulação teórica não apenas questionava os limites do estruturalismo, como também fortalecia o movimento da chamada “Revolução Cognitiva”, aqui já apresentada, recolocando o estudo da mente no centro das investigações científicas sobre a linguagem. Sua obra seminal, *Syntactic Structures* (1957), marcou o início de um novo paradigma na linguística, deslocando o foco da observação do comportamento linguístico para a investigação das estruturas internas que possibilitam a criatividade da linguagem.

Chomsky postulou a existência de uma Gramática Universal (GU), uma faculdade inata e geneticamente programada, inerente à espécie humana. A hipótese sobre a GU se altera ao longo desses mais de 50 anos de gerativismo, mantendo sempre que não se trata de um conjunto de regras específicas de uma língua, mas sim um sistema de princípios e parâmetros abstratos, como apresentado em sua obra *Lectures on Government and Binding* (1981), que servem como arcabouço para a aquisição de qualquer língua natural. No modelo mais atual do Minimalismo, a universalidade ganha contornos distintos, porque são respostas a pressões do sistema cognitivo. Essa faculdade inata opera como um dispositivo de aquisição da linguagem (LAD, do inglês – *Language Acquisition Device*), cuja função primordial é viabilizar a aquisição de uma língua qualquer a partir da interação com a experiência linguística, ou *input* linguístico. Nesse processo, a experiência, embora essencial como fonte de dados, não é vista como o fator criador do conhecimento linguístico, mas sim como o insumo necessário que o LAD utiliza para ativar e parametrizar as estruturas abstratas da GU. Desse modo, a exposição a um conjunto limitado e, por vezes, imperfeito de sentenças permite que o dispositivo gere um sistema de conhecimento linguístico completo e robusto.

Assim, explica-se um dos argumentos centrais de Chomsky em favor da GU, o princípio da pobreza de estímulo. Esse conceito refere-se à observação de que a linguagem disponível no ambiente – o *input* linguístico recebido pelas crianças – é frequentemente incompleto, limitado e imperfeito, contendo ambiguidades, erros e lacunas que não poderiam, por si só, explicar a aquisição completa da gramática de uma língua. Apesar dessas limitações, além da complexidade de uma língua, as crianças, excetos os casos patológicos, adquirem competências linguísticas complexas de forma rápida e uniforme, sem esforço e sem necessidade de ensino, produzindo e compreendendo sentenças que jamais ouviram anteriormente. Esse princípio é um grande elemento quando estamos comparando a produção e geração de linguagem por humanos com modelos de LLMs, pois, como veremos mais adiante, esses modelos precisam de um vasto campo de dados linguísticos para poderem gerar uma sentença. No entanto, por princípio, a mente humana poderia ser uma supermáquina indutora. O que parece enfraquecer a hipótese é precisamente que as mentes humanas não são super indutoras. Se fossem deveriam se comportar como as máquinas, mas não parece que seja esse o caso, como veremos.

Assim, para Chomsky, esse fenômeno só pode ser explicado pela existência de estruturas linguísticas inatas, fornecidas pela GU, que permitem ao indivíduo extrapolar regras e gerar uma linguagem plenamente funcional a partir de um *input* incompleto. A GU forneceria as restrições e os princípios operacionais necessários para que a criança construa a gramática a partir da exposição limitada a dados primários. Por exemplo, a regra de interpretação de movimento do verbo em sentenças encaixadas no inglês em (1a) e a interpretação do pronome 'ele' na sentença em (1b):

(1) a. The boy who is tall is sick.
b. Ele disse que João está doente.

Esses são casos clássicos na literatura. A pobreza de estímulo, portanto, reforça a ideia de que a criatividade e a produtividade linguísticas não decorrem exclusivamente da experiência ambiental, mas surgem da interação entre o *input* linguístico e a GU. Dessa forma, a experiência não é a causa da linguagem, mas a catalisadora que transforma o conhecimento inato da mente humana em competência na língua específica do ambiente.

Central à teoria gerativa da época foi a distinção entre competência e performance. A competência refere-se ao conhecimento tácito e idealizado que o falante-ouvinte possui da gramática de sua língua – um sistema de regras internalizado que permite a geração e interpretação de sentenças. A performance, por sua vez, diz respeito ao uso efetivo da linguagem em situações concretas, que pode ser influenciada por fatores extralinguísticos, como lapsos de memória, erros de fala ou interrupções contextuais. Para Chomsky, a linguística teórica deve priorizar o estudo da competência. Essa distinção foi abandonada nas propostas mais recentes, embora ela tenha sido fundamental para consolidar a linguística como uma ciência voltada para a investigação das propriedades intrínsecas da mente/cérebro humano, e não apenas dos produtos observáveis das formas de interação linguística. A criatividade e a recursividade linguística são, nesse sentido, manifestações da competência: a capacidade inata de gerar e compreender um número infinito de sentenças a partir de um conjunto finito de regras. A abordagem chomskyana, ao fundamentar a Faculdade da Linguagem, abriu um vasto campo de pesquisa, ao direcionar o estudo para os mecanismos cognitivos subjacentes que permitem a

aquisição, a produção e a compreensão de estruturas linguísticas complexas. Permitiu compreendermos com muito mais acuidade as propriedades das línguas humanas.

Assim, a psicolinguística passou a estudar experimentalmente como esses mecanismos cognitivos operam durante o uso da linguagem. Pesquisas sobre tempo de reação e memória de trabalho buscam compreender como o cérebro organiza e interpreta sentenças complexas em tempo real. Por exemplo, experimentos que analisam a compreensão de sentenças ambíguas como (2):

(2) “O policial viu o homem com o telescópio”

investigam como os falantes resolvem ambiguidades sintáticas, o que fornece evidências sobre os processos cognitivos envolvidos na interpretação linguística. Indicam também como a mente processa as estruturas sintáticas, privilegiando as mais simples estruturalmente, por exemplo, e permite verificar se há dependência contextual, entre outras questões.

Nesse sentido, a teoria de Chomsky revela um dos grandes diferenciais da produção humana, central a reflexão e comparação com mecanismos estatísticos, aqui os chamados LLMs, ausentes de uma faculdade inata, criatividade ou qualquer relação com contextos além dos fornecidos de imediato pelo usuário. Desde o início de sua obra, incluindo as mais recentes (Chomsky, 2023), o autor ainda pontua que os sistemas de IA, em contraste, são modelos estatísticos de comportamento verbal, carentes dos princípios recursivos e da sensibilidade ao contexto que caracterizam a cognição humana.

TURING E A MÁQUINA

A complexidade da inteligência humana, capaz de aprender e gerar línguas naturais diversas, sempre representou um desafio para a ciência. Desse modo, a replicação artificial do fenômeno da produção da linguagem só se tornou concebível a partir do trabalho de Alan Turing (1912-1954), que percebeu ser possível usar a máquina de escrever como inspiração para criar o seu modelo de máquina automática, com base no que ele chamou de “configuração atual da máquina”, ou seja, a máquina poderia em qualquer determinado momento estar em número finito de possíveis combinações. Esses estudos, que levaram Turing a criar sua própria

máquina, foram essenciais para a fundação do que hoje conhecemos como IA e, consequentemente, os modelos de LLMs analisados neste trabalho.

Em seu artigo de 1936, "*On Computable Numbers, with an Application to the Entscheidungs problem*", Turing introduziu o conceito da Máquina de Turing. Esse modelo matemático abstrato serviu como o alicerce teórico para os computadores modernos. A máquina hipotética demonstrou que o raciocínio humano, por mais complexo que seja, poderia ser simulado por um sistema que opera com base em um conjunto finito de regras lógicas. Essa ideia, que parecia abstrata na época, foi fundamental para estabelecer que a inteligência humana – ou parte dela, como defende esse trabalho – poderia ser replicada por meio de um sistema lógico-formal.

A contribuição de Turing para a linguagem e a IA atingiu seu ponto mais alto com a publicação de "*Computing Machinery and Intelligence*" em 1950. Neste trabalho, ele não apenas propôs a possibilidade de máquinas "pensarem", mas também realizou uma redefinição fundamental do problema filosófico que a questão implicava. A pergunta inicial, "poderiam as máquinas pensar?", foi prontamente reconhecida por Turing como excessivamente complexa e insolúvel, dada a subjetividade inerente ao conceito de "pensamento". Ele argumentou que a tentativa de definir o pensamento seria uma busca infrutífera. Assim, sugeriu que, em vez de focar na consciência interna de uma máquina (algo inobservável), o foco deveria ser transferido para o seu comportamento externo, sua capacidade de responder.

A genialidade de sua abordagem foi a substituição de uma questão ontológica por uma questão pragmática, dando origem ao Teste de Turing. No teste, uma pessoa pode estar se comunicando por texto com duas entidades, uma humana e uma máquina, sem saber qual é qual. Se o interrogador não fosse capaz de distinguir a máquina do ser humano, a máquina passaria no teste. A relevância dessa metodologia para o campo da IA é monumental. Turing estabeleceu um critério objetivo para avaliar a inteligência de uma máquina por meio da sua capacidade de usar a linguagem de forma convincente.

Em essência, a visão de Turing abriu caminho para o desenvolvimento desses sistemas. Ele forneceu tanto a abstração teórica para que o pensamento pudesse ser codificado (a Máquina de Turing) quanto a metodologia para que a inteligência pudesse ser avaliada (o Teste de Turing). Suas ideias, portanto, não serviram apenas de inspiração, mas formaram a base sólida sobre a qual opera todo o campo da ciência da computação.

No entanto, com o surgimento de LLMs com o Chat GPT, Bert e Gemini, o próprio conceito do teste parece ter se tornado obsoleto. Modelos modernos já alcançaram uma capacidade de conversação tão avançada que, em ambientes controlados, conseguem passar no teste. Um estudo publicado na PNAS em 2023, por exemplo, mostrou que o ChatGPT-4 foi considerado humano em metade das vezes (Mei *et al*, 2024).

Desse modo, o foco contemporâneo concentra-se no aprimoramento da confiabilidade e robustez dos sistemas, com ênfase na mitigação de alucinações, na expansão de capacidades multimodais e de especialização, que permitam a aplicação prática e segura dessas ferramentas em contextos complexos do mundo real, sendo que a grande pressão e a influência para sofisticação desses sistemas tem um viés completamente econômico. A métrica de sucesso, portanto, já não é a capacidade de se passar por um humano, mas sim a de funcionar como uma ferramenta útil e confiável, e para compreendermos melhor o funcionamento da linguagem humana, na medida em que mostram os limites da máquina.

PROCESSAMENTO DE LINGUAGEM NATURAL (PLN)

O Processamento de Linguagem Natural (PLN) é um campo multidisciplinar, sua função operacional se dedica a treinar computadores para compreender, interpretar e gerar qualquer língua natural. Sua evolução, no entanto, não seguiu uma linha linear, mas se desenvolveu através de mudanças teóricas e metodológicas de paradigmas que moldaram sua arquitetura e aplicação. Contudo, é importante destacar que, quando falamos em compreensão e interpretação por modelos de IA, não nos referimos ao fenômeno da linguagem executados por humanos, que correlacionam as sentenças produzidas a experiências vívidas, mas sim a um sistema lógico-computacional. O PLN permite que computadores e dispositivos digitais reconheçam, entendam e gerem texto e fala, apenas metaforicamente (Shanahan, 2024, ao combinar linguística computacional com modelagem estatística e aprendizado de máquina, como veremos a seguir.

As bases conceituais do PLN foram firmemente estabelecidas por pensadores como Alan Turing, como apresentado na seção anterior. Essa proposta deslocou o foco para um critério comportamental, validando a ideia de que a inteligência de uma máquina poderia ser avaliada pela sua capacidade de se comunicar de forma

indistinguível de um ser humano (Turing, 1950). A partir dessa premissa, o campo do PLN ganhou um objetivo tangível.

Posteriormente, outro ponto que analisamos nos capítulos iniciais também exerceu influência profunda no PLN. Noam Chomsky, com sua teoria da Gramática Gerativa, estabeleceu que a linguagem humana possui uma estrutura inata e universal (Chomsky, 1957). Essa perspectiva influenciou os primeiros sistemas de PLN, que tentavam replicar a linguagem por meio de um mapeamento formal de regras gramaticais.

Inicialmente, o caminho para alcançar esse objetivo foi dominado pela abordagem simbólica. Sistemas como o ELIZA (Weizenbaum, 1966), considerado um dos primeiros exemplos de interação em linguagem natural entre humanos e máquinas, o qual simulava uma psicoterapeuta chamada Eliza, respondendo às entradas do usuário por meio de simples regras de correspondência de padrões, foram construídos sobre um vasto conjunto de regras gramaticais, codificadas manualmente por especialistas. Essa metodologia, contudo, logo demonstrou suas limitações diante da complexidade, ambiguidade e irregularidade inerentes às línguas naturais. A criação de regras se mostrou insustentável para cobrir todas as nuances linguísticas, e a falta de escalabilidade dessa abordagem levou a uma busca por novos paradigmas (Manning & Schütze, 1999).

A virada decisiva ocorreu com a emergência do paradigma estatístico e do aprendizado de máquina. Abandonando as regras manuais, os pesquisadores passaram a treinar modelos em grandes volumes de dados textuais, permitindo que os próprios algoritmos aprendessem a probabilidade de certas palavras e sequências ocorrerem.

No entanto, foi o advento do aprendizado profundo (*deep learning*) que elevou o PLN a um novo patamar. O desenvolvimento de arquiteturas como as Redes Neurais Recorrentes (RNNs) e, mais notavelmente, os *Transformers*, permitiu que os modelos capturassem dependências contextuais de forma sem precedentes (Kamath, Liu & Whitaker, 2024). A arquitetura *Transformer* e seu inovador mecanismo de atenção (Vaswani *et al*, 2017) superaram as limitações dos modelos anteriores ao permitir que o sistema pesasse a importância de cada palavra em relação a todas as outras em uma sentença de comando enviada pelo usuário, independentemente da distância. Isso viabilizou o treinamento de modelos massivos de linguagem, como

GPT e BERT aqui analisados, sistemas de LLMs que hoje são a base de diversas aplicações.

LARGE LANGUAGE MODELS (LLM)

Por trás da aparente “criatividade e compreensão” dos LLMs está uma operação matemática executada em uma escala monumental: a previsão probabilística. Esses sistemas operam com um princípio fundamental, prever a próxima palavra mais provável em uma sequência, com base em padrões complexos apreendidos com vastos conjuntos de dados. O funcionamento dos LLMs pode ser entendido como um processo em três atos principais: tokenização, processamento por camadas de atenção e geração auto-regressiva (Jurafsky & Martin, 2024).

O processo de segmentação em unidades processáveis, conhecido como tokenização, é a etapa fundamental do PLN, na qual o texto de entrada é segmentado em unidades discretas e quantificáveis, denominadas *tokens*. Esse processo converte a linguagem humana, que é complexa, em uma representação estruturada que pode ser manipulada por modelos computacionais. Os *tokens* podem corresponder a palavras completas, morfemas, ou caracteres individuais, e cada um é subsequentemente mapeado para um identificador numérico único, criando o vocabulário base do modelo (Jurafsky & Martin, 2024).

Após a tokenização, os *tokens* numéricos são transformados em *embeddings*, que são vetores de alta dimensão. Esses vetores são a base para a representação semântica, pois codificam a informação contextual e relacional de cada *token* em um espaço vetorial contínuo. A proximidade geométrica entre os vetores nesse espaço reflete a similaridade semântica entre os *tokens*, capturando o que o modelo aprendeu durante o treinamento em grandes *corpora* de texto. A qualidade e a capacidade desses *embeddings* de capturar significado são cruciais para o desempenho do modelo em tarefas de linguagem, uma vez que eles fornecem a representação de entrada para as camadas subsequentes da rede neural (Scalabrin, 2024).

A camada de atenção, uma inovação central na arquitetura dos *Transformers*, permite que o modelo compute um contexto ponderado para cada *token* na sequência. Ao contrário de modelos sequenciais tradicionais que processam o texto em uma ordem fixa, a atenção calcula a relevância de cada *token* em relação a todos os outros *tokens* na sequência de entrada, atribuindo-lhes pesos de atenção. Esse

mecanismo confere ao modelo a capacidade de ponderar a importância de *tokens* distantes no texto, resolvendo dependências de longo alcance e permitindo uma compreensão mais holística do contexto. O cálculo da atenção é essencial para a performance em tarefas complexas, pois permite ao modelo focar nos *tokens* mais informativos para a tarefa atual. (Scalabrin, 2024).

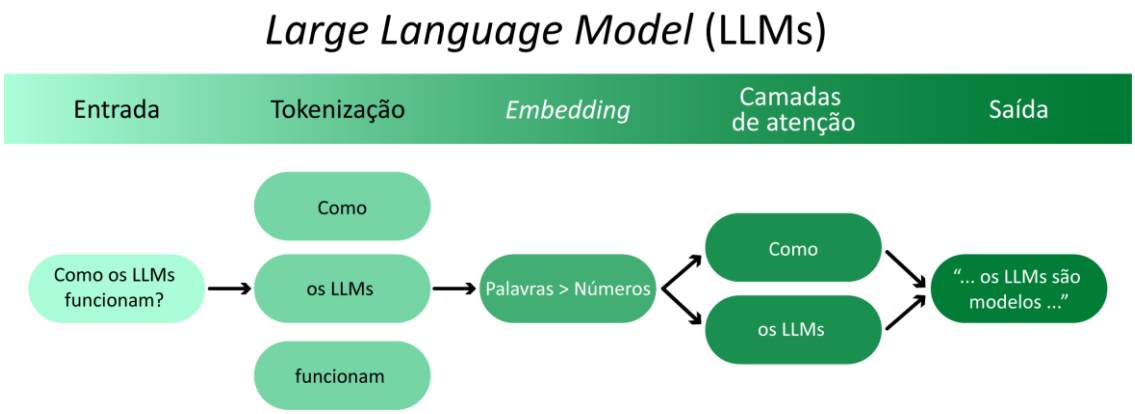
A etapa final do processamento envolve a geração de texto, que ocorre através da previsão do próximo *token* na sequência. A rede neural, utilizando a representação contextual enriquecida pelas camadas de atenção, calcula uma distribuição de probabilidade sobre todo o seu vocabulário para determinar o *token* mais provável a seguir. O modelo então seleciona um *token* a partir dessa distribuição e o anexa à sequência. Esse ciclo de processamento, em que cada novo *token* gerado se torna parte do contexto para a próxima previsão, continua até que um *token* de parada seja gerado ou a sequência atinja o seu comprimento máximo. Esse processo iterativo é o que habilita a capacidade de geração de texto coerente e fluído em LLMs (Scalabrin, 2024).

As camadas de atenção e as redes neurais dos transformadores processam simultaneamente todos os *tokens* da entrada, computando as correlações e o contexto fornecido no *prompt* – comando enviado pelo usuário – de forma instantânea. Essa dualidade é a chave para a notável eficiência dos LLMs, permitindo-lhes gerar texto de maneira fluida e coerente (Vaswani *et al*, 2017).

Em sua essência, o processamento de um LLM combina uma natureza seriada na geração com uma execução simultânea em sua arquitetura interna. Embora a geração de texto proceda de forma sequencial, com cada novo *token* sendo predito e adicionado à sequência um de cada vez, o cálculo que torna essa previsão possível opera em um paralelismo massivo.

A imagem abaixo representa, de forma simplificada, o esquema do processamento de linguagem em LLMs:

Figura 1 - Representação do sistema de processamento dos LLMs



Fonte: Elaboração própria

É relevante para a presente análise pontuar que a etapa de tokenização constitui um processo fundamentalmente distinto da decomposição morfológica realizada pelo processamento linguístico humano. Enquanto o cérebro humano opera através de princípios linguísticos que segmentam palavras em constituintes mórficos portadores de significado (como radicais e afixos), os sistemas de tokenização em LLMs implementam algoritmos estatísticos que fragmentam o texto em unidades baseadas exclusivamente em frequência de ocorrência nos dados de treinamento.

Enquanto uma criança, nos primeiros anos de aprendizado, é capaz de produzir uma sentença como “eu sabo”, fazendo correlação com outras conjugações de verbo como “eu falo”, já demonstrando um rico repertório e conhecimento morfológico da língua, a abordagem dos LLMs resulta em *tokens* que são essencialmente unidades probabilísticas, desprovidas de qualquer noção de morfologia ou gramática. Conforme demonstrado pela literatura especializada, modelos como GPT não "compreendem" que ‘funcionam’ decompõe-se em ‘funcion-’ (radical) + ‘-am’ (marca de 3 pessoa plural), mas sim aprendem, através de padrões correlacionais em trilhões de *tokens*, que a sequência ‘funcion’ + ‘am’ ocorrem frequentemente em contextos específicos, gerando representações vetoriais que capturam essas associações estatísticas (Goldberg, 2019). Esta limitação inerente aos métodos estatísticos de tokenização revela-se particularmente problemática diante de neologismos ou construções linguísticas não observadas durante o treinamento, expondo o caráter associativo do processamento linguístico nas arquiteturas contemporâneas de LLMs.

Em Amouyal, Meltzer-Asscher e Berant (2024), os autores propuseram a seguinte aplicação metodológica: utilizar LLMs como ferramenta de "pré-teste de

plausibilidade" psicolinguística. O estudo demonstrou que os LLMs, quando instruídos a julgar a plausibilidade de sentenças, apresentam uma significativa correlação com os julgamentos humanos, superando modelos linguísticos anteriores. No entanto, os próprios autores ressaltam que o sucesso do modelo está condicionado ao seu treinamento com dados de conhecimento linguístico humano. "*A possible theoretical explanation for this phenomenon is that the outputs of LMs can be viewed as an average over multiple samples, since pretraining is done on texts from many authors*" (Amouyal, Meltzer-Asscher Berant, 2024, p. 9).

Se o objetivo inicial de Turing era a imitação comportamental da linguagem, os modelos atuais demonstraram uma capacidade de processamento que expandiu drasticamente o que se considerava possível, embora a distinção entre a reprodução de padrões e a compreensão genuína ainda permaneça como uma questão central para o futuro da pesquisa em IA.

Ao comparar esses mecanismos biológicos com os LLMs, fica evidente que, embora os modelos computacionais possam simular a performance linguística – com exceção de alguns fenômenos, como apresentado durante todo o texto, e das alucinações, criação de respostas falsas apenas pela necessidade de responder positivamente aos *prompts* –, com alguma precisão, a natureza de sua "compreensão" difere fundamentalmente daquela observada no cérebro/mente humano.

O estudo de Linzen e Leonard (2018) emprega fenômenos gramaticais específicos para comparar diretamente o processamento sintático em modelos neurais e no cérebro humano. Os autores focam no fenômeno de concordância sujeito-verbo. Os resultados revelaram um paralelo superficial: ambos os sistemas cometeram erros de concordância. No entanto, a análise detalhada dos padrões de erro expôs uma divergência fundamental. Enquanto os erros humanos foram sistemáticos e influenciados pela estrutura hierárquica da sentença, os erros do modelo estudado seguiram um padrão marcadamente diferente, sendo frequentemente baseados em associações superficiais e estatísticas locais presentes nos dados de treinamento.

Apesar de suas saídas exibirem uma organização superficial que se assemelha à sintaxe e à semântica humanas, o modelo em si não parece manipular estruturas sintáticas hierárquicas ou conteúdos semânticos genuínos. O que emerge como uma "sentença gramatical" ou um "conceito coerente" é, na realidade, o

resultado da atualização de pesos em uma rede neural para prever a sequência de palavras mais provável dado um contexto, sem conseguir se referir a elementos fora do contexto imediato.

Portanto, a "sintaxe", a "semântica" e a “morfologia” observadas são resultados de um processo puramente estatístico, não operando com representações simbólicas ou com a intencionalidade que fundamenta a linguagem humana.

A ausência de uma base biológica experiencial e a dependência de vastos bancos de dados para o aprendizado, contrastam com a eficiência e a integração multimodal do processamento cerebral. A capacidade do cérebro/mente de processar a linguagem, integrando-a a outras funções cognitivas a partir de dados limitados, sugere uma abordagem de representação fundamentalmente distinta daquela utilizada pelos sistemas de LLMs.

REFLEXÕES COMPARATIVAS ENTRE O PROCESSAMENTO LINGUÍSTICO DE HUMANOS E LLMS

Uma das distinções mais marcantes entre o processamento linguístico humano e o dos LLMs reside na quantidade de dados necessária para a aquisição da competência linguística. O cérebro humano demonstra uma eficiência notável na aprendizagem da linguagem, sendo capaz de adquirir estruturas gramaticais complexas e um vocabulário extenso a partir de um conjunto de dados relativamente reduzido, como vimos na teoria da pobreza de estímulo (Chomsky, 1965), com uma idade bem tenra e em pouco tempo. A quantidade de palavras as quais a criança é exposta na infância se torna insignificante quando comparada aos trilhões de *tokens* utilizados no treinamento de LLMs como GPT-4. Esta disparidade evidencia diferenças profundas nos mecanismos subjacentes de aprendizagem: enquanto os humanos operam com uma faculdade da linguagem e um limitado *corpus*, os LLMs operam através de aprendizado estatístico com um *corpus* quase imensurável.

O estudo de Houghton, Kazanina e Sukumaran (2023) oferece uma contribuição fundamental ao debate ao situar o desempenho dos LLMs no contexto histórico das teorias do processamento linguístico humano. Os autores partem da famosa crítica de Chomsky aos modelos de n-gramas – mecanismo utilizado em IAs para capturar dependências locais e padrões estatísticos em dados sequenciais, como texto ou fala –, que os considerava inadequados para explicar a produtividade

e a criatividade da linguagem humana, por serem fundamentalmente limitados. A investigação central do artigo consiste em testar se os LLMs modernos, enquanto sistemas também baseados em previsão probabilística, superam essas limitações conceituais ou se constituem meramente em n-gramas de escala astronômica. Através da análise de padrões linguísticos específicos, como dependências de longo alcance – que ocorre quando dois ou mais elementos em uma sentença estão gramaticalmente conectados, mas separados por uma distância significativa, com outras palavras, frases ou até cláusulas entre eles – e a generalização sistemática de estruturas sintáticas, os autores demonstram que os LLMs de fato exibem capacidades que transcendem as restrições dos modelos puramente baseados em estatísticas de superfície locais. No entanto, eles argumentam de forma crucial que esta superação não equivale a uma replicação do mecanismo cognitivo humano. Em vez disso, os LLMs parecem operar através de um "funcionalismo computacional", no qual alcançam resultados similares por meio de uma arquitetura baseada em princípios subjacentes radicalmente diferentes. Esta conclusão sustenta uma distinção essencial entre a competência linguística humana, internalizada e guiada por princípios gramaticais, e a performance estatística dos LLMs, que emerge de sua exposição massiva a dados.

A capacidade de gerar verdadeira novidade linguística constitui outro domínio de nítido contraste entre os dois sistemas. A criatividade humana manifesta-se não apenas na produção de sentenças originais, mas na capacidade de criar novos significados – como quando passamos a chamar de nuvem não somente aquele amontoado branco no céu, mas o espaço virtual no qual podemos salvar nossos dados –, metáforas inéditas e formas linguísticas anteriormente inexistentes, como ‘namorido’. Esta criatividade emerge de uma interface rica entre a linguagem e outras experiências, como emoção, experiência corporal e conhecimento de mundo, que permitem ao falante reconhecer “palavras” como coisas reais do mundo e adaptá-las às suas necessidades. Evidências empíricas demonstram, por exemplo, que lesões no giro fusiforme – tradicionalmente associado ao reconhecimento visual de faces e objetos – comprometem significativamente a compreensão de palavras referentes a entidades visuais, revelando a interdependência entre representações lexicais e sistemas perceptivos (Binder *et al*., 2009).

O sistema cognitivo humano ancora a linguagem em modalidades perceptivas primárias. O significado da palavra ‘quente’ é inextricavelmente ligado a experiências

térmicas somatossensoriais, até o momento em que um novo significado possa ser atribuído a essa palavra e, quando adicionado ao contexto, resultará em uma nova ligação para os humanos. Essa ligação às experiências sensório-motoras contrasta radicalmente com a arquitetura dos LLMs, que operam sobre representações puramente simbólicas, desprovidas de experiências que fundamentam a linguagem humana.

Um cenário que exemplifica a natureza essencialmente derivativa e combinatória dos LLMs pode ser verificado na geração de informações solicitadas explicitamente nos *prompts*. Por exemplo, se solicitado a encontrar uma citação de “biólogo francês chamado XYZ”, um LLM pode compor, de forma coerente e estilisticamente adequada, uma passagem que inclui um pesquisador inexistente, atribuída a uma instituição real e a uma publicação periódica plausível. Esta produção não emana de um ato intencional de referência ou de um conhecimento factual, mas sim da exploração de padrões estatísticos profundos internalizados durante o treinamento. O modelo combina probabilisticamente elementos lexicais e sintáticos associados ao gênero "discurso acadêmico" – como a estrutura "[Autor, (Ano)] argumenta que..." e o uso de terminologia especializada – para gerar uma sequência textual que simula autoria e fundamentação. Consequentemente, a "criatividade" exibida manifesta-se como uma representação superficial da norma, destituída de experiência, consciência e domínio das máximas conversacionais (Grice, 1975) que caracteriza a comunicação humana genuinamente criativa e intencional.

Pode-se argumentar que a capacidade de gerar genuína novidade conceitual é uma função direta da semântica – o componente da linguagem que lida com a referência ao mundo, ao falante e a construção de sentido. Esses aspectos parecem inerentes ao diálogo humano, uma vez que um LLM é fundamentalmente um sistema reativo, projetado para responder a *prompts* e completar tarefas. Por sua vez, um interlocutor humano pode, legitimamente, responder a uma pergunta com um "não sei" seguido de um silêncio constrangedor, mudar abruptamente de assunto por cansaço ou tédio, produzir implicações ou mesmo contestar a premissa da pergunta. Os LLMs, contudo, são otimizados para a utilidade e continuidade, tornando respostas como "isso não me interessa" ou a recusa em engajar com um tópico sem uma razão explícita estatisticamente improvável. Estes operam fundamentalmente em um nível distribucional e estatístico, mapeando relações formais entre *tokens* sem qualquer ancoragem ontológica em um "estado real das coisas". A sua produção, por mais

coerente que seja sintaticamente, é essencialmente uma recombinação de padrões linguísticos internalizados, destituída de um modelo mental do mundo que confira veracidade, crença ou uma compreensão causal aos símbolos que manipula. A "fala" de um LLM é um serviço de processamento de informação, que buscará sempre gerar uma resposta utilitária, o que também torna o sistema mais propenso às alucinações e à geração de respostas falsas, numa tentativa de responder positivamente a todos os comandos. Portanto, a grande lacuna dos LLMs reside precisamente no domínio semântico: eles carecem da capacidade de formar representações internas com conteúdo verídico sobre a realidade, o que os impede de realizar o salto criativo que caracteriza a cognição humana autêntica, permanecendo, assim, engenhosos simulacros de compreensão.

Um LLM é, em sua base, um mecanismo de otimização. Seu "objetivo" primordial, embutido em sua arquitetura matemática, é sempre o mesmo: minimizar a surpresa estatística. Ele é treinado para prever a próxima palavra mais provável e, em operação, é configurado (via *prompting* e ajustes) para gerar a resposta mais "útil" ou "provável" dentro de um contexto apresentado pelo usuário. Toda a sua existência computacional é uma busca por eficiência na tarefa que lhe foi apresentada. Ele não "pensa", como o próprio Turing já nos apresentou, se quer responder; ele executa o processo para o qual foi desenhado. Isso os torna fundamentalmente diferentes. E, ao estudar essa diferença, nós não estamos apenas entendendo as limitações da IA; estamos, na verdade, definindo e redescobrindo o que é mais singular na experiência humana.

CONCLUSÃO

O presente trabalho buscou elucidar, através de uma breve revisão bibliográfica, teórica e experimental, a questão central: “como o cérebro/mente humana e os sistemas de LLMs se comportam no processo de interpretação/decodificação e geração de estruturas ou sequências linguísticas, quais atividades são realizadas e, mais crucialmente, se há e quais as razões para as distinções observáveis entre eles”. A análise da literatura confirma uma disparidade ontológica fundamental: os LLMs operam por meio de inferência estatística em representações vetoriais, aprendendo padrões de associação em um *corpus* de dados massivos, enquanto a cognição linguística humana é situada, baseando-se em

experiências sensoriais-motoras, consciência e intencionalidade (Lakoff & Johnson, 1999). A literatura em filosofia aponta os “qualia”, os sensíveis, como aquilo que só entes sencientes possuem; diferenciando das máquinas. Nas versões mais atuais do gerativismo, no modelo minimalista, a Faculdade da Linguagem é um sistema computacional otimizado que resulta como uma resposta à pressão da cognição que é guiada por suas experiências sensoriais. Como bem lembra Chomsky, em vários de seus escritos, a Faculdade da Linguagem é única aos humanos. Assim, na dicotomia *mind/brain*, o cérebro corresponderia à capacidade biológica de gerar e interpretar linguagem e a mente corresponderia à consciência corpórea que guia a interpretação dos significados e as escolhas das respostas.

Já o trabalho de Turing trouxe à luz um ponto importante para a discussão dos LLMs. A máquina não “pensa”, ela tenta imitar um sistema com base em ocorrência e probabilidade. Então, qual é o ponto? Se as máquinas já passaram no teste, ou estão muito próximas disso, por que a comunidade científica não celebra uma vitória definitiva? A resposta, creio eu, reside no fato de que a relevância do teste não está mais em sua aplicação, mas em sua limitação. O Teste de Turing avalia a capacidade de imitação, não de compreensão ou de consciência ou de sensibilidade. Um LLM pode replicar padrões de fala de forma convincente, mas operando em um nível puramente distribucional e estatístico, sem a ancoragem ontológica no mundo real que caracteriza a linguagem humana. O trabalho de Houghton Kazanina e Sukumaran (2023) nos lembra que essa imitação bem-sucedida é um caso de 'funcionalismo computacional' – a máquina alcança resultados similares por meio de uma arquitetura subjacente diferente da humana. Mesmo assim, e talvez por isso mesmo, se coloca uma questão ética, se elas se comportam como um humano e podem ser utilizadas para enganar os humanos, de diferentes maneiras.

O foco, portanto, mudou. A busca agora não é por um sistema que apenas se pareça inteligente, mas por um que demonstre raciocínio, criatividade e a capacidade de resolver problemas de forma verdadeiramente autônoma, diminuindo a densidade de alucinações e melhorando o desempenho em interpretar fenômenos linguísticos complexos – esses, até onde pudemos observar, ligados à diferentes aspectos da produção, ou seja, informações não ditas nos *prompts*, que carecem de interpretação contextualizada.

Assim, a questão de criar uma máquina genuinamente criativa e imprevisível como o ser humano permanece em aberto. A imprevisibilidade humana – mudar de

assunto, declarar desinteresse, ficar em silêncio – não é um defeito, mas a manifestação de um sistema com desejos, cansaços e uma história de vida própria. A capacidade de produzir respostas inovadoras, de criar novos significados e de ter *insights* inesperados é um atributo central da inteligência humana. Em contraste, as máquinas, mesmo as mais avançadas, são inerentemente determinísticas e utilitárias; suas respostas, por mais sofisticadas que pareçam, são o resultado da reprodução e combinação de padrões depreendidos a partir de um conjunto de dados fornecidos. A imprevisibilidade que se manifesta em sistemas como os LLMs é, na realidade, uma consequência da vasta quantidade de dados processados e da complexidade de seus parâmetros, e não uma genuína capacidade de pensamento espontâneo.

Portanto, a natureza estritamente utilitária das máquinas não é um detalhe técnico a ser superado com modelos maiores ou mais dados. É uma consequência do fato de que elas não têm uma vida, um corpo, uma representação e a singularidade do ‘eu’. Carecem dos fundamentos necessários não apenas para a intencionalidade genuína, mas também para a eficiência semântica e a economia de aprendizado que definem a inteligência humana.

E é por isso que é possível descrever com considerável precisão a arquitetura operacional dos LLMs: uma sequência determinística de transformações matemáticas e otimizações estatísticas aplicadas a um espaço vetorial de *tokens*. No entanto, o mesmo grau de clareza não se aplica ao processamento linguístico humano. Este opera a partir de uma base neurobiológica cuja emergência de propriedades como a consciência, a intencionalidade e a integração sensorial permanece em parte um mistério científico. Podemos, assim, especificar o *design* da máquina, mas não podemos replicar a cognição humana a um modelo computacional inequívoco. A proficiência estatística de um LLM em recombinar os dados seria, portanto, uma paródia da cognição, pois esta é, antes de tudo, prospectiva e criadora. O abismo entre ambos não é apenas técnico, mas epistemológico: os LLMs são um artefato cujo funcionamento é, em última instância, conhecível porque foi construído; a inteligência humana, um sistema natural cuja complexidade integral ainda não parece possível de replicar.

Ao final deste percurso, somos confrontados pela reflexão fundamental que orienta o futuro da IA: as coisas que as máquinas ainda não fazem podem ser ensinadas? Ou há algo intrinsecamente humano, uma centelha de ser no mundo, que coordena essas coisas e que uma máquina nunca poderá aprender?

Apesar do terreno vasto e das inúmeras áreas ainda por explorar nesta interface entre linguística e IA, um questionamento crítico se impõe: será este o caminho ideal? Num contexto de recursos energéticos finitos, investir no desenvolvimento e no estudo aprofundado de sistemas que replicam competências que os seres humanos já dominam com maestria, merece uma reflexão mais aprofundada. O avanço dos LLMs é, inegavelmente, parte do presente e do passado recente da humanidade, e muito certamente o futuro. A investigação comparativa entre a mente humana e os LLMs é válida e produtiva, oferecendo *insights* tanto para a neurociência quanto para a ciência da computação. No entanto, a corrida para aperfeiçoá-los e compreender suas operações internas consome quantidades colossais de recursos energéticos, financeiros e intelectuais, enquanto muito pouco recurso é destinado à pesquisa básica sobre as línguas humanas, sobre as quais ainda sabemos pouco.

A fascinação pela replicabilidade da nossa própria cognição corre o risco de negligenciar o estudo direto da própria fonte: a diversidade das línguas humanas. Enquanto bilhões são investidos em fazer com que os LLMs gerem textos ligeiramente mais coerentes ou "alucinem" menos, milhares de línguas humanas enfrentam a extinção iminente, levando consigo visões únicas, sistemas de conhecimento e manifestações da nossa capacidade de criação e adequação (Evans & Levinson, 2009).

A imensa variedade de línguas, muitas das quais permanecem por documentar e estudar, mostra-se a manifestação esperada de uma faculdade flexível e adaptativa, cuja principal universalidade reside precisamente na sua capacidade de gerar uma pluralidade irredutível de soluções para o desafio da comunicação humana. Cada uma destas línguas é um experimento natural, milenar, em inteligência e adaptação. Ignorar este laboratório vivo em favor de um foco quase exclusivo em sistemas artificiais que espelham, predominantemente, um punhado de línguas hegemônicas, é uma oportunidade epistemológica perdida de proporções monumentais. O caminho ideal não seria o abandono desta frente, mas um reequilíbrio estratégico.

A advertência final do neurocientista Miguel Nicolelis (Altman, 2020) ecoa como um veredito sobre um futuro puramente guiado pela IA: “Se tudo o que você vai fazer daqui para frente é baseado em um banco de dados do que já foi feito, você não tem futuro”. O verdadeiro futuro – da ciência, da arte e da própria humanidade – deverá residir na curadoria e recombinação do passado ou na capacidade de criar, a

partir do nada, da experiência vivida, um significado que nunca antes existiu? É nesse domínio, que reside a fronteira final e talvez intransponível para a máquina.

REFERÊNCIAS

ALTMAN, B. **Inteligência Artificial**: tudo o que você precisa saber - Miguel Nicolelis - programa 20 minutos. Entrevista com Miguel Nicolelis. Youtube, 12 jun. 2023. Disponível em: https://www.youtube.com/live/pb4b4_MlNwo?si=ohm_HpPhyid8FIf8. Acesso em: 4 nov. 2025.

AMOUYAL, S. J.; MELTZER-ASSCHER, A.; BERANT, J. **Large Language Models for Psycholinguistic Plausibility Pretesting**. arXiv, , 8 fev. 2024. Disponível em: http://arxiv.org/abs/2402.05455. Acesso em: 1 ago. 2025.

BINDER, J. R.; DESAI, R. H.; GRAVES, W. W.; CONANT, L. L. **Where is the semantic system**? A critical review and meta-analysis of 120 functional neuroimaging studies. Cerebral Cortex, Maryland, 2009. Disponível em: https://pubmed.ncbi.nlm.nih.gov/19329570/. Acesso em: 6 jul. 2925.

BLOOMFIELD, L. **Language**. 1. ed. New York: Henry Holt and Company, 1933.

CHOMSKY, N. **A review of B. F. Skinner's Verbal Behavior**. Language, Baltimore, v. 35, n. 1, p. 26-58, 1959. Disponível em: https://chomsky.info/1967____/. Acesso em: 6 jul. 2025.

CHOMSKY, N. **Aspects of the Theory of Syntax**. Cambridge, MA: MIT Press, 1965.

CHOMSKY, N. **Lectures on Government and Binding**. The Pisa Lectures. Dordrecht: Foris Publications, 1981.

CHOMSKY, N. Noam Chomsky: The False Promise of CHATGPT. **The New York Times**. Disponível em: https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html. Acesso em: 29 set. 2025.

CHOMSKY, N. **Syntactic structures**. The Hague: Mouton & Co., 1957.

CORREA, L.M.S. Língua e Cognição: Antes e depois da revolução cognitiva. *In*: GUIMARÃES, E (ed.). **Introdução às ciências da linguagem**: linguagem história e conhecimento. Campinas: Editoras pontes, 2006.

DESCARTES, R. **Discurso do Método**. Tradução de Maria Ermantina Galvão. 3. ed. São Paulo: Martins Fontes, 2001.

EVANS, N.; LEVINSON, S. C. **The myth of language universals**: Language diversity and its importance for cognitive science. Cambridge University Press, v. 32, n. 5, p, 2009. Disponível em: https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/myth-of-language-universals-language-diversity-and-its-importance-for-cognitive-science/25D362A6566FCA4F51054D1C41104654. Acesso em: 20 out. 2025.

FITCH, W. T.; HAUSER, M.; CHOMSKY, N. **The evolution of the language faculty: clarifications and implications**. Cognition, n. 97, 2005.

FREESTONE, M.; SANTU, S. K. K. **Word Embeddings Revisited**: Do LLMs Offer Something New? arXiv, , 2 mar. 2024. Disponível em: <http://arxiv.org/abs/2402.11094>. Acesso em: 17 set. 2024

GOLDBERG, Y. **Assessing BERT's Syntactic Abilities**. 2019. arXiv. Disponível em: https://arxiv.org/abs/1901.05287. Acesso em: 12 jul. 2025.

HOUGHTON, C.; KAZANINA, N.; SUKUMARAN, P. **Beyond the limitations of any imaginable mechanism**: large language models and psycholinguistics. arXiv, 28 fev. 2023. Disponível em: http://arxiv.org/abs/2303.00077. Acesso em: 1 ago. 2025.

HUMBOLDT, W. V. **Sobre o Pensamento e a Linguagem**: Escritos de Humboldt. Tradução e apresentação de Antonio Ianni Segatto. São Paulo: Editora da Universidade de São Paulo (Edusp), 2009.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing.** 3. ed. Stanford Edu. 2024. Disponível em: https://web.stanford.edu/~jurafsky/slp3/. Acesso em: 17 out. 2025.

KAMATH, U.; LIU, J.; WHITAKER, J. **Deep learning for NLP and speech recognition**. Cham, Switzerland; Springer, 2024. Disponível em: https://link.springer.com/book/10.1007/978-3-030-14596-5. Acesso em: 17 out. 2025.

LAKOFF, G.; JOHNSON, M. **Philosophy in the Flesh**: The Embodied Mind and its Challenge to Western Thought. Basic Books, 1999.

LINZEN, T; LEONARD, B. **Distinct patterns of syntactic agreement errors in recurrent networks and humans**. arXiv, , 18 jul. 2018. Disponível em: http://arxiv.org/abs/1807.06882. Acesso em: 24 maio 2025.

MAIA, M. (Org.). **Psicolinguística, psicolinguísticas**: uma introdução. São Paulo: Editora Contexto, 2015.

MANNING, C. D.; SCHÜTZE, H. **Foundations of Statistical Natural Language Processing**. Cambridge, MA: MIT Press, 1999.

MEI, Q. *et al.* **A bilingual parallel corpus with discourse annotations for Chinese-English machine translation.** Proceedings of the National Academy of Sciences, [s. l.], v. 121, n. 16, p. 1-10, 2024. Disponível em: https://www.pnas.org/doi/10.1073/pnas.2313925121. Acesso em: 28 jun. 2025.

MORAES, L. C. *et al.* **Análise de ambiguidade linguística em modelos de linguagem de grande escala (LLMs)**. arXiv e-prints, 1 abr. 2024. Disponível em: https://ui.adsabs.harvard.edu/abs/2024arXiv240416653D. Acesso em: 20 nov. 2025.

PINKER, S. **The Language Instinct**: How the Mind Creates Language. New York: William Morrow and Company, 1994.

SCALABRIN, E. E. **Introdução ao processamento de linguagem natural**. FERNANDES, D. S. A. (org.) *et al*. Goiânia: Cegraf/UFG, 2024. E-book (174 p.) ISBN: 978-85-495-1038-9. Disponível em: https://portaldelivros.ufg.br/index.php/cegrafufg/catalog/view/649/621/2570. Acesso em 25 fev. 2025.

SCHLENKER, P.; CHEMERO, A.; FULLER, J.; GAUNET, F.; SEGUIN, M.; CASTRES, L. **Formal monkey linguistics**. Theoretical Linguistics, Berlim, v. 42, n. 1-2, p. 1-198, 2016. Disponível em: https://doi.org/10.1515/tl-2016-0001. Acesso em: 24 maio. 2025.

SKINNER, B. F. **Verbal Behavior**. New York: Appleton-Century-Crofts, 1957.

TRAXLER, M. J.; GERNSBACHER, M. A. (ed.). **Handbook of Psycholinguistics.** 2 ed. London, Academic Press. 2006.

TURING, A. M. **Computing machinery and intelligence**. Mind, London, v. 59, n. 236, p. 433-460, Oct. 1950.

TURING, A. M. **On Computable Numbers, with an Application to the Entscheidungs problem**. Proceedings of the London Mathematical Society, [s. l.], s2-43, n. 1, 1936. Disponível em: https://www.cs.virginia.edu/~robins/Turing_Paper_1936.pdf. Acesso em: 27 jun. 2025.

VASWANI, A. *et al.* **Attention is all you need**. Proceedings: 4 Dec. 2017. Disponível em: https://dl.acm.org/doi/10.5555/3295222.3295349. Acesso em: 23 out. 2025.

WEIZENBAUM, J. ELIZA: a computer program for the study of natural language communication between man and machine. *In*: JOINT COMPUTER CONFERENCE, Spring, 1966, Washington. **Proceedings**. New York: ACM, 1966. p. 1-15. Disponível em: https://dl.acm.org/doi/10.1145/365153.365168. Acesso em: 27 jun. 2025.

LISTA DE SIGLAS

LLM	Large Language Models
IA	Inteligência Artificial
LAD	Language Acquisition Device
GU	Gramática Universal
PLN	Processamento de Linguagem Natural
RNN	Rede Neural Recorrente