

DOSSIÊ

Narrativas de IA: tendências da
produção audiovisual



V. 15 – N. 1 - jan./abr. 2024

ISSN: 2179-1465 / <https://www.revistageminis.ufscar.br>
DOI: <https://doi.org/10.14244/2179-1465.RG.2023v14i3p44-53>

O PROBLEMA DA (FALTA DE) MEMÓRIA EM NARRATIVAS GERADAS POR INTELIGÊNCIA ARTIFICIAL

THE PROBLEM OF (LACK OF) MEMORY IN NARRATIVES GENERATED BY
ARTIFICIAL INTELLIGENCE

EL PROBLEMA DE (FALTA DE) MEMORIA EN NARRATIVAS GENERADAS POR
INTELENCIA ARTIFICIAL

Fabio Cardoso

Universidade Estadual Paulista (UNESP)
ORCID: <https://orcid.org/0009-0005-8932-8647>
Bauru, SP, Brasil

Recebido: 20/01/2024 / Aprovado: 16/04/2024

Como citar: CARDOSO, F. O Problema da (falta de) Memória em Narrativas Geradas Por Inteligência Artificial.
Revista GEMInIS, v. 15, n. 1, p. 44–53, 2024.

Direito autoral: Sob os termos da Licença Creative Commons-Atribuição 3.0 Internacional.



RESUMO

A evolução das ferramentas de geração de texto por inteligência artificial abre a possibilidade do uso criativo da tecnologia para criação de roteiros audiovisuais. Modelos de linguagem como o ChatGPT são capazes de, com uma entrada do usuário, gerar conteúdo artístico complexo. Isso viabiliza o surgimento de uma nova modalidade de conteúdo audiovisual: a geração contínua de roteiros. No entanto, dificuldades técnicas inerente do modelo atual de criação desses softwares limitam a eficácia e abrangência dessa forma de criação. Este artigo busca investigar o problema mais significativo desta limitação, que é a falta de memória dos algoritmos de LLM.

Palavras-chave: Large Language Models. Memória. Roteiros Audiovisuais.

ABSTRACT

The evolution of artificial intelligence text generation tools opens up the possibility of creatively using technology for the creation of audiovisual scripts. Language models like ChatGPT are capable of generating complex artistic content with user input. This enables the emergence of a new form of audiovisual content: continuous script generation. However, inherent technical difficulties in the current model of creating such software limit the effectiveness and scope of this form of creation. This article aims to investigate the most significant problem of this limitation, which is the lack of memory in LLM algorithms.

Keywords: Large Language Models. Memory. Audiovisual Screenplay.

RESUMEN

La evolución de las herramientas de generación de texto por inteligencia artificial abre la posibilidad del uso creativo de la tecnología para la creación de guiones audiovisuales. Modelos de lenguaje como ChatGPT son capaces de generar contenido artístico complejo con la entrada del usuario. Esto permite el surgimiento de una nueva modalidad de contenido audiovisual: la generación continua de guiones. Sin embargo, las dificultades técnicas inherentes al modelo actual de creación de este tipo de software limitan la eficacia y alcance de esta forma de creación. Este artículo tiene como objetivo investigar el problema más significativo de esta limitación, que es la falta de memoria en los algoritmos de LLM.

Palabras Clave: Large Language Models. Memoria. Guiones Audiovisuales.

INTRODUÇÃO

A evolução das ferramentas de geração de texto por inteligência artificial, capazes de gerar conteúdo complexo através de treinamento prévio, abre a possibilidade do uso criativo da tecnologia para criação de roteiros audiovisuais. Modelos de linguagem como o *ChatGPT* e o *Google Gemini* são capazes de, através uma entrada adequada do usuário, gerar conteúdo artístico complexo e verossímil, o que viabiliza o surgimento de uma nova modalidade de conteúdo audiovisual: a geração contínua de roteiros sequenciais “infinitos”. Se devidamente instruído para tal, um modelo de linguagem moderno pode, em tese, gerar perpetuamente resultados coesos, com a próxima interação sempre levando em conta o contexto da saída anterior. No entanto, dificuldades técnicas inerentes ao modelo atual destes softwares limitam a eficácia e abrangência dessa forma de criação. Este artigo busca investigar o problema mais significativo destas limitações, que é a falta de memória dos algoritmos de *LLM* guiados por inteligência artificial.

Inteligência artificial é um termo guarda-chuva que abrange uma diversidade de tecnologias que trabalham juntas, possibilitando que um algoritmo busque reproduzir a cognição humana, percebendo o ambiente e contexto em que está, e tomando decisões levando estes fatores em consideração (RUSSEL E NORVIG, 1995, p. 8). Desde a criação dos computadores de uso geral, no período pós Segunda Guerra Mundial, a comunicação entre humanos e máquinas se dá, prioritariamente, através do texto escrito. Computadores trabalham em uma interface binária, e adaptações para introduzir uma melhor comunicação entre humanos e máquinas são necessárias. Para isso, programadores criaram softwares que tentavam viabilizar a produção de linguagem natural pelos computadores, que é o modo com o qual humanos comunicam-se.

Uma das soluções encontradas pelos programadores para facilitar a comunicação entre humanos e máquinas foram os “modelos de linguagem”, que são softwares que conseguem receber entradas de texto e devolver saídas de texto pertinentes, levando em conta o contexto de uma conversação. Os modelos computacionais de linguagem surgiram em 1966, quando o *ELIZA*, primeiro algoritmo a usar técnicas de modelos de linguagem para emular a conversação entre humanos, foi criado por Joseph Weizenbaum (1966). No início do século, com a introdução dos algoritmos guiados por inteligência artificial, os algoritmos ganham robustez e eficácia, e, juntamente com a evolução da infraestrutura que os move, tornou possível o surgimento dos Grandes Modelos de Linguagem, ou *Large Language Models (LLM)*.

Large Language Models, como o *ChatGPT* e o *Google Gemini*, apropriam-se de conceitos de inteligência artificial e possuem a capacidade de processar e gerar linguagem natural de forma complexa e sofisticada. Os modelos artificiais de linguagem disponíveis atualmente possuem

características de desenvolvimento comuns, como a utilização de mecanismos de atenção (que atuam na compreensão do contexto da entrada do usuário), e a característica de serem pré-treinados: Todo o conhecimento do modelo de linguagem é processado durante seu treinamento inicial, abastecido com um vasto corpus de informações, de bilhões, ou até trilhões, de *tokens*¹.

O acrônimo *GPT*, em inglês, significa “*Generative Pre-trained Transformer*”: “*Generated*” advém da sua natureza generativa, “*Pre-trained*” indica que o modelo de linguagem é fechado e toda sua informação é dada pelo banco de dados inicial, e “*Transformer*” é o nome da tecnologia de rede neural que define as relações contextuais entre os tokens de texto. Por ser pré-treinada e ter suas respostas limitadas ao conjunto de dados apresentados inicialmente pelo seu desenvolvedor, os *LLM* criados sob este modelo só podem utilizar, para geração de sua saída, o conjunto de dados de seu próprio treinamento. Isso não é um problema para um aplicativo de IA generativa que lide com dados genéricos e relativamente imutáveis, como os geradores de faces humanas como o site “*This person does not exist*”², mas pode ser um problema para um *LLM* que se proponha a gerar conteúdo textual contextualizado, devido a informações temporalmente defasadas. Além da possível defasagem temporal, a falta de memória recente de um modelo de *LLM* como o *ChatGPT* causa outro problema, quando a geração de conteúdo sequencial e contextualizado é necessária: a falta de memória recente.

Por ser pré-treinado, não há um mecanismo que simule a característica humana de memória recente para um modelo de linguagem treinado por IA. Ao devolver uma resposta para o usuário, o modelo encerra seu processamento e aguarda por uma nova entrada, que não possui vínculo algum com a entrada anterior. O modelo não leva em conta nenhuma informação da pergunta anterior, não aprende com ela e não traz nenhuma informação sobre o contexto.

Todavia, uma das características marcantes da verossimilhança nos diálogos com estes *LLM* são as respostas sequenciais, onde a pergunta seguinte é respondida como se os diálogos anteriores estivessem em um fluxo de informação bidirecional, com um contexto preservado. Este fator contextual é dado pela técnica da repetição de entrada: tanto as próximas entradas, quanto as respostas anteriores dadas pelo modelo, são repassadas para a máquina na pergunta seguinte. Esta entrada composta é omitida da interface do chat, para passar uma impressão fidedigna de diálogo. Conforme uma conversa evolui, e possivelmente vários assuntos são citados, o contexto se modifica. Em um diálogo longo entre dois humanos, há uma constante troca de informações, que modifica e reprograma o contexto da conversa, agindo no processo de enunciação, moldando a forma como construímos e

¹ Os tokens são conjuntos de caracteres que formam a base do texto. Eles são produzidos por um processo chamado tokenização, que divide o texto em partes menores seguindo certas regras, como espaços, pontuação e caracteres especiais. Embora os tokens muitas vezes representem palavras, nem sempre é o caso.

² Disponível em <https://thispersondoesnotexist.com>

interpretamos a linguagem. Essa memória, composta pelas experiências, conhecimentos e emoções mais recentes, atua como um filtro dinâmico que influencia a seleção de signos, a organização do discurso e a construção do sentido.

Ao enunciarmos algo, nos deparamos com um universo gigantesco de signos linguísticos à disposição, fornecidos por toda as nossas vivências, experiências e memórias que adquirimos ao longo da vida. Escolher quais destes signos usar é um processo que não se limita a um mero exercício de seleção dentre um repertório disponível em nosso “banco de dados” do cérebro, mas também leva em conta fatores momentâneos, como o local do diálogo, das nossas experiências, do nosso estado emocional e da análise destes fatores que fazemos do enunciatário do nosso discurso. Em um diálogo com uma IA, faz parte do “contrato fiduciário” (GREIMAS e COURTÉS, 2008, p. 86) entre o usuário e a máquina geradora o fato de que a máquina não possui os fatores emocionais característicos de humanos: os que se propõem a participar dessas interlocuções já sabem, de antemão, que a conversa não sofrerá a influência de emoções do lado maquínico. Porém, dois fatores que são comuns em diálogos entre humanos e não ocorrem nos diálogos entre humanos e máquinas, são simulados pela estrutura conversacional dos *LLM* e deixam a impressão de estarem presentes, mas não o estão: A memória recente e o aprendizado.

Esse simulacro do contexto realizado pelo aplicativo de inteligência artificial é efetivo para conversações menores, mas padece de um problema estrutural que só é revelado com usos específicos, com sequências longas de geração, ou geração ininterrupta: há um limite do tamanho máximo da entrada do modelo, dado pela capacidade do algoritmo e da máquina cujo algoritmo está rodando. O modelo mais avançado hoje da *OpenAI*³, o *GPT 4*, possui um limite máximo de 128 mil *tokens* para sua entrada. Esta limitação técnica dos *LLM* exige que todo o contexto anterior de uma conversa seja transmitido na entrada que precede a próxima geração, possibilitando à máquina a chance de entender todo o passado do diálogo e entregar uma resposta dentro de um diálogo natural para o usuário.

O USO DE IA NOS ROTEIROS AUDIOVISUAIS

O uso crescente de recursos de inteligência artificial em produtos e serviços é um fenômeno global, e tem um impacto significativo em várias áreas do conhecimento humano, dentre elas, a comunicação (KAPLAN e HAENLEIN, 2019, p. 15-16). A integração da inteligência artificial no domínio da comunicação está induzindo uma reconfiguração paradigmática tanto na gênese quanto

³ A OpenAI é um laboratório privado de pesquisa em IA e uma empresa que tem como objetivo desenvolver a IA e direcioná-la de maneiras que "beneficiem toda a humanidade" (descrição dada pela própria empresa).

na disseminação de informação, e as implicações e aplicações dessa intersecção abrangem várias áreas desta ciência, como a produção de conteúdo automatizada, personalização e segmentação automática da audiência, análise sentimental e avaliação de feedback, monitoramento midiático, análise de tendências, publicidade direcionada, melhorias na acessibilidade, dentre outras.

A inserção de recursos de inteligência artificial na indústria audiovisual aumentou a eficiência de processos, abriu novas possibilidades criativas, contribuiu para a fidelização e engajamento da sua audiência e ampliou o acesso a ferramentas antes restritas a grandes produções, diminuindo a barreira financeira para a produção de conteúdo de áudio e vídeo (DU e HAN, 2021, p. 2)

Recursos de inteligência artificial também podem atuar nas etapas criativas do processo artístico. As capacidades dos *LLM* de conduzirem conversas contínuas e potencialmente infinitas, com base em entradas anteriores, permitem a criação de diálogos extensos, que podem ser utilizados para a geração de roteiros audiovisuais. Essa característica, presente nos *LLM* modernos, viabiliza a criação de roteiros de ficção com um fluxo contínuo: os próximos diálogos e desenvolvimentos das cenas são gerados com base nos diálogos anteriores, criando assim uma obra audiovisual teoricamente infinita.

Nessa premissa de geração contínua de texto, foi criada a sitcom gerada por inteligência artificial "*Nothing, Forever*", uma produção audiovisual americana gerada de forma contínua e transmitida ao vivo pela internet. Esta foi concebida por Skyler Hartle, engenheiro de software, e Brian Habersberger, físico de polímeros, ambos sem expertise prévia na área audiovisual. Ambos são membros do Mismatch Media, um coletivo de arte experimental. A sitcom é uma derivação da série *Seinfeld*, que foi transmitida na televisão americana de 1989 a 1998, e pertencia ao gênero de narrativa conhecido como "slice of life", que retrata fragmentos da vida cotidiana dos personagens, muitas vezes enfatizando experiências mundanas ou rotineiras. O criador e protagonista da série original, Jerry Seinfeld, a descreveu como "um show sobre o nada", destacando o humor derivado de situações simples da vida cotidiana. *Nothing, Forever* adapta esse conceito de humor sobre o "nada" (*Nothing*) e o combina com o elemento de infinita continuidade (*Forever*).

A série possui duas temporadas, sendo que a primeira foi transmitida durante três meses, de 14 de dezembro de 2022 a 6 de fevereiro de 2023, quando foi interrompida devido a uma suspensão no serviço de streaming ao vivo *Twitch*, que hospedava o programa. Durante um monólogo de *stand-up comedy* realizado pelo personagem Larry, uma característica também herdada de *Seinfeld*, onde o protagonista é um comediante e trechos de seu show eram apresentados na série, foram proferidas falas consideradas inadequadas pela plataforma.

Embora as capacidades contextuais dos *LLM*, dadas pelas subseqüentes repetições do contexto nas entradas posteriores, sejam notáveis, sua aplicação na geração contínua de conteúdo encontra limitações conforme a progressão da narrativa. Os personagens, imersos em um fluxo constante de criação de conteúdo, geram situações que impactam a si mesmos, outros personagens (inclusive aqueles fora de cena) e o ambiente circundante. Em alguns casos, tais influências podem se estender além do microcosmo da interação dos personagens, como exemplificado pela retirada do programa do ar devido a comentários considerados inadequados feitos pelo protagonista. É esperado (aqui sob o escrutínio da semiótica discursiva de linha francesa), dentro do contrato fiduciário entre enunciador e enunciatário, que a continuidade da história seja mantida através da absorção dessas situações, com os personagens moldando-se e evoluindo em resposta a elas.

Em roteiros audiovisuais convencionais, tais ações são decididas pelo autor, que considera as implicações das ações ficcionais dos personagens e define suas conseqüências na narrativa. Por exemplo, se uma personagem fica grávida no início da trama e a história se estende por um período superior a nove meses, é necessário que o desfecho dessa gravidez seja abordado de alguma forma, seja o nascimento do bebê, o aborto, o desaparecimento da personagem ou outra explicação plausível. No entanto, no caso de um roteiro gerado por uma ferramenta de *LLM*, as inferências são limitadas aos textos de treinamento inicial e ao possível contexto fornecido na entrada da geração de texto, o que gera um problema significativo à medida que a história se desenvolve. Após um certo período, torna-se inviável introduzir todo o histórico do personagem, cenários, eventos e personagens secundários em cada entrada de texto, dada a limitação de memória disponível.

Nothing, Forever exemplifica várias dessas limitações. Os fãs registram os diálogos e eventos da série em uma wiki, contextualizando a história ao longo do tempo. No entanto, algumas informações geradas pela IA para o personagem principal, como sua habilidade de falar espanhol, não são armazenadas no *LLM* e acabam sendo “esquecidas” nas interações posteriores. Essas inconsistências narrativas são notáveis na comédia situacional, prejudicando a imersão do espectador. É de se esperar que esses problemas de memória narrativa se intensifiquem à medida que a história avança, gerando mais inconsistências causadas pelas limitações de memória de um *LLM*. Embora avanços na tecnologia possam eventualmente mitigar essas limitações, atualmente elas representam um obstáculo significativo para a performance narrativa de produtos criativos gerados por inteligência artificial.

A estrutura narrativa de *Nothing, Forever* leva em consideração essas limitações. O programa é episódico, com períodos específicos de geração de diálogos em blocos com começo, meio e fim. Esses blocos são intercalados com a simulação de um guia de programação fictício de outros canais

fictícios. Cada episódio tem uma duração definida, dentro da qual as interações entre personagens são mantidas em um contexto relativo. No entanto, após o término de um episódio, todas as informações geradas durante aquele período são esquecidas, e um novo bloco é gerado com as especificações iniciais. Embora a narrativa pareça contínua devido à temática do programa e à impossibilidade prática de acompanhar o show por longos períodos, cada bloco possui um contexto fechado que não é levado adiante nos blocos subsequentes.

No estado atual da tecnologia, as limitações de memória de um *LLM* não são um problema para a criação roteiros audiovisuais mais curtos; a geração de um roteiro de um longa-metragem, por exemplo, não é afetada por essa situação. O filme possui um começo, um meio e um fim – aqui, não se tratando do conceito narrativo de início e fim, mas, sim, de seu significado temporal: ele começa, e acaba. Obras derivadas e sequências podem ser posteriormente realizadas, mas a obra original não é afetada.

Produtos audiovisuais fragmentados, como uma telenovela, onde (comumente) a obra está indo ao ar enquanto novos capítulos são escritos, podem ser afetados pelo feedback do usuário, tendo seu percurso narrativo modificado de acordo com a recepção do público. Quem faz este ajuste é o autor, que é humano, e dotado de memória. Ele pode processar, avaliar e julgar, dentro da sua visão de autor da história, como as considerações e impressões do público que acompanha a narrativa devem afetar o produto. O único modo de se transmitir estes feedbacks a um *LLM* é declará-los nas entradas subsequentes de geração de novos roteiros, e, necessariamente, tais novas entradas devem conter, além destes dados, todos os dados anteriores de composição de personagens, acontecimentos, e demais fatos relevantes para a narrativa.

Em um exemplo: solicita-se a um *LLM* que crie um roteiro dos primeiros episódios sobre uma família que possui vários animais. O algoritmo cria três episódios, e, no terceiro, o roteiro aponta que houve uma festa na casa dessa família, e o gato de estimação fugiu pela janela, e não retornou. Estes episódios foram ao ar, e vários espectadores manifestaram sentimentos sobre a perda do gato, que, esperam, tenha abalado a família que se importa com os animais de estimação que tem. Porém, ao solicitar ao *LLM* que criasse novos episódios para a série, como o algoritmo não possui memória, os próximos capítulos mostram a família vivendo novas aventuras, com o gato vivendo entre os outros animais sem nenhuma menção ao animal ter fugido anteriormente.

Para evitar este problema, os autores da série poderiam ter declarado, durante o pedido inicial para a criação de outros episódios, que o gato da família estava desaparecido, que o *LLM* criaria histórias levando este fator em conta. No entanto, há um limite no tamanho da entrada de texto de um *LLM*, que pode influenciar na capacidade do software de absorver o contexto: de um modo não

otimizado, o usuário poderia abastecer a entrada seguinte com todo o conteúdo da geração anterior, situando a máquina do que ela mesma gerou anteriormente. Aqui temos um problema matemático: A cada entrada nova, soma-se a entrada anterior. Se um roteiro de um episódio fictício desta série for composto de 10 páginas de texto (com 500 palavras em cada página), cada entrada terá 5 mil palavras. A seguinte conterá 10 mil, e atinge-se o limite do GPT-4 no décimo segundo episódio. Pode-se usar estratégias para diminuir esta limitação – como pedir a própria IA para resumir a entrada anterior, mantendo apenas os aspectos chave – contudo, eventualmente, o limite de tokens do sistema seria atingido e a série teria sua narrativa prejudicada.

CONSIDERAÇÕES FINAIS

Estes problemas técnicos na geração de conteúdo por algoritmos de inteligência artificial reduzem, no momento, o potencial desta tecnologia, obrigando os criadores que optem por esta modalidade de criação a adaptarem seus recursos criativos às limitações da plataforma. Já há iniciativas, como a do Google Gemini 1.5, que expandem estes limites a níveis muito elevados (no caso, mais de 1 milhão de tokens), logo, espera-se, em breve que esta limitação seja minimizada, ou resolvida.

A introdução de narrativas audiovisuais geradas autonomamente representa um marco disruptivo no panorama da indústria audiovisual contemporânea. Essas narrativas, fundamentadas na capacidade de algoritmos de produzir conteúdo dinâmico e infinitamente variável, oferecem uma nova perspectiva sobre os processos de criação, distribuição e consumo de mídia. Ao desafiar os paradigmas estabelecidos, essa tecnologia propõe uma reconfiguração da lógica de produção e da interação entre criadores e público. Do ponto de vista da produção, a geração processual de conteúdo audiovisual implica uma revisão fundamental das metodologias tradicionais, permitindo aos criadores empregarem algoritmos para desenvolver narrativas adaptativas que respondam às interações do público ou a conjuntos de dados específicos. Essa abordagem não apenas aumenta a eficiência ao reduzir os tempos de produção e os custos associados, mas também promove a criação de obras mais personalizadas e adaptativas, capazes de atender às expectativas e aos interesses de uma audiência diversificada.

No entanto, essa evolução também levanta questões pertinentes acerca da autoria, da propriedade intelectual e dos direitos autorais. A geração de conteúdo por meio de algoritmos desafia as noções convencionais de criação e originalidade, exigindo uma reavaliação das bases legais e éticas que regem a produção cultural. A introdução deste modelo de criação de narrativas pode ser responsável por uma transformação profunda na indústria audiovisual, impactando todos os aspectos

relacionados à produção e ao consumo de conteúdo. Embora essa tecnologia ofereça oportunidades sem precedentes para a inovação e a personalização, ela também apresenta desafios significativos que exigem uma reflexão cuidadosa sobre as implicações técnicas, éticas, culturais e legais. As limitações técnicas impostas pela estrutura algorítmica dos *LLM* tende a se dissiparem com o tempo, e inovações tecnológicas são uma constante nesta área, pavimentando um futuro onde as narrativas contínuas geradas por inteligência artificial poderão – tecnicamente – serem viáveis, sem restrições. Considerações que ultrapassam o nível analítico e movem questões mais filosóficas e sociológicas da abordagem, como a repercussão no quadro de trabalho do cenário audiovisual, questões autorais, e até o sentido artístico de narrativas robotizadas certamente suscitarão discussões acadêmicas pertinentes.

REFERÊNCIAS BIBLIOGRÁFICAS

DU, W.; HAN, Q. **Research on Application of Artificial Intelligence in Movie Industry**. 2021 International Conference on Image, Video Processing, and Artificial Intelligence. Shanghai: [s.n.]. 2021.

GREIMAS, A. J.; COURTÉS, J. **Dicionário de Semiótica**. São Paulo: Contexto, 2008.

KAPLAN, A.; HAENLEIN, M. **Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations and implications of artificial intelligence**. Business Horizons, 2019.

NOTHING, Forever. Direção: S. HARTLE e B. HABERSBERGER. [S.l.]: Mismatch Media. 2023.

WEIZENBAUM, J. **ELIZA - a computer program for the study of natural language communication between man and machine**. Communications of the ACM, jan. 1966.

RUSSEL, S.; NORVIG, P. **Artificial Intelligence: A Modern Approach**. 4. ed. [S.l.]: Pearson, 1995.

Informações sobre o Artigo

Resultado de projeto de pesquisa, de dissertação, tese: Tese.

Fontes de financiamento: Não se aplica

Apresentação anterior: Não se aplica

Agradecimentos/Contribuições adicionais: Não se aplica

Fabio Cardoso

Doutor em Comunicação pela UNESP – Universidade Estadual Paulista.

E-mail: fabio.cardoso@unesp.br

ORCID: 0009-0005-8932-8647