**Como eu crio Agentes de IA com Memória Longa e Humanizada - Tutorial N8n & ZEP**

**Transcrito por** [**TurboScribe.ai**](https://turboscribe.ai/pt/?ref=docx_export_upsell)**.** [**Atualize para Ilimitado**](https://turboscribe.ai/pt/subscribed?ref=docx_export_upsell) **para remover esta mensagem.**

E aí pessoal, Will Pires aqui, e nesse vídeo eu vou apresentar pra vocês como eu construo agentes de IA com a memória ultralonga pra deixar a conversa muito mais personalizada, centrada no cliente e voltada a gerar valor pro cliente com alguns insights valiosos, isso deixando mais humanizada a conversa entre o usuário final e o agente de IA. Eu bati muito a cabeça pra desenvolver esse framework de conversas humanizadas que eu utilizo hoje na criação dos meus agentes de IA, e eu vou nesse vídeo aqui apresentar pra vocês algumas coisas que eu bati a cabeça no passado, algumas sugestões pra você agregar mais valor ainda aos seus projetos de IA, se você tá criando agentes de IA, e em cima de tudo isso, um tutorial passo a passo da ferramenta que eu tô utilizando hoje pra criação dessa memória longa, eu tô utilizando a ferramenta da Zap AI, eu não sou afiliado deles, mas eu tô recomendando ela pra criação desses agentes de IA, ok? Então beleza pessoal, chegando aqui, primeiro de tudo, antes da gente pular pra um tutorial com a plataforma, eu nomeei de Paradoxo Memoryshed justamente esses conceitos que a gente acaba quebrando cabeça na hora de criar agentes de IA personalizados e humanizados. Bem, se o resultado que você quer é um agente de IA humanizado, que traga insights pro usuário final, que gere valor pro usuário final e seja centrado no usuário final, você tem que parar de ser centrado no produto, então algumas coisas ferramentais você tem que superar, a primeira delas é o prompt, basicamente, sendo totalmente direto ao ponto, a primeira delas é o prompt.

A maioria das pessoas gasta muito tempo, energia e gasta muita cabeça mesmo na hora de criar um prompt pros agentes de IA, seja por dois motivos, ou porque não consegue ser direto e objetivo no prompt pro agente IA, ou não tem dimensão do que o agente de IA pode fazer ou não. Por exemplo, o cenário real da minha agência, Maven AI Solutions, eu tô desenvolvendo projetos de IA há 18 meses e eu recebo cotações de projetos todos os dias, as pessoas me procuram pra fazer projetos de IA, que alguns deles são meio alucinógenos, alguns deles são fora de órbita, como por exemplo, já recebi solicitação de pessoas que querem automatizar a máquina de lavar louça ou então querem automatizar a portaria do prédio utilizando um agente de IA no WhatsApp, o que acaba sendo um pouco fora de órbita de fazer e quando as pessoas tentam fazer isso, querem colocar no prompt da IA pra fazer essas coisas loucas, então imagina, você tá criando um agente de IA nessa plataforma tipo N8n e você fala eu quero que quando eu falar oi você ligar a minha máquina de lavar louça, esse tipo de coisa você vai gastar o seu prompt, cada letra que você coloca no seu prompt são quatro tokens que você tá consumindo no seu agente de IA e cada token é como se fosse uma ficha que você tá apostando pro seu projeto dar certo. Quanto mais fichas você aposta, maior a chance de dar errado, te falando assim, quanto mais tokens você coloca pro seu agente IA digerir de uma vez, menor a chance dele trazer um resultado sólido pra você.

Você pode até achar que se você der muita instrução a coisa pode dar muito bom e pode de fato dar muito bom, mas a maior chance é de dar errado, porque você gastou várias fichas em vão e você vai ter um retorno pequeno e não vai dar lucro nessa aposta de fichas. Para pra pensar no Cassino de Las Vegas. Então bem, sobre a parte de você ter consciência no seu prompt de IA pra realidade, trazendo agora exemplos da realidade, foram esses exemplos alucinógenos que eu trouxe pra você fora de órbita, existem os exemplos de pessoas que querem que a IA faça realmente, por exemplo, ah eu quero que você automatize a minha CRM no HubSpot, ok, você vai dar essa instrução pro seu agente de IA, mas quem vai fazer a automação no HubSpot é a própria ferramenta de automação, o N8n, e não seu agente de IA, então você não precisa ir tão longe assim no prompt do seu agente de IA pra fazer esse tipo de coisa e eu vejo muita gente, engenheiros de automação pecando nisso daí, colocando a IA como se fosse uma Jarvis, uma assistente do Homem de Ferro, Iron Man, que você dá uma ordem e ela vai fazer XYZ sem nem pensar, sendo que não é assim, a maioria dos desenhos de processos são feitos com as ferramentas de automação e o agente de IA é 20% do projeto.

Então bem, isso daqui é o primeiro paradigma que eu queria quebrar, que o erro geralmente tá no prompt. O segundo é que o agente deve buscar as mensagens antigas com o cliente, mas isso são só palavras armazenadas no banco, isso daqui é meio que uma punchline, uma citação, vamos falar aqui do Well, ficou meio difícil de escrever aqui, mas é uma citação minha, eu não quero viralizar com essa citação, isso daqui não é pra fazer um corte, mas se você parar pra pensar logicamente, esses bancos de memória que a gente utiliza, tipo Postgres, Superbase, e a gente quer armazenar as conversas do nosso usuário com o nosso agente de IA utilizando essas ferramentas, só que pra fazer as consultas, meio que a gente só tá armazenando a palavra, as palavras que foram utilizadas, meio que sem contexto nenhum, então esse banco de memória transforma as nossas palavras em números e tudo vira um jogo de matemática, essa palavra aqui é 3, e se uma palavra for parecida, a gente vai fazer a análise semântica, esse tipo de coisa, então o valor da palavra fica muito quadrado, se o único destino dela é um número, não é a equação, que é o que acaba acontecendo com esses bancos. E por consequência, uma coisa ligada com a outra, se o usuário não fizer uma consulta com palavras muito semelhantes ao que ele tinha falado anteriormente, a busca no banco irá falhar.

A maneira que esses bancos fazem esse tipo de consulta é assim que funciona, a busca é semântica, então beleza, uma palavra que seja parecida com a outra vai trazer aquela primeira palavra, então você fala, a raça do meu cachorro é Golden Retriever, e depois você pergunta pra IA, qual a raça do meu cão, ou qual a raça do meu bicho, agora se você falar, qual a raça do meu melhor amigo, sendo que você lá atrás, lá atrás, lá atrás, há muito tempo atrás falou que seu melhor amigo é o seu cachorro, pode ser que se o seu banco de dados for muito, muito, muito eficiente, o seu prompt for eficiente, você vai dar certo, mas é uma série de estrelas que tem que se alinhar pra dar certo, o banco de dados vai te permitir isso, mas só vai dar certo, você só vai conseguir extrair se você seguiu muito bem os fundamentos, e o que eu quero justamente é passar um atalho pra você fazer esse fundamento dar certo, ao invés de você colocar todas as suas fichas no prompt, ou colocar todas as suas fichas no memory chat, você tem um framework, um passo a passo pra fazer isso daqui dar certo, e assim as suas chances vão aumentar, deixar o prompt mais curto, e buscar um banco de chat memory que faça mais sentido pro seu agente de IA. Então no final, as pessoas querem que o seu agente de IA seja humanizado, mas não se importam que o agente de IA tenha um processo de memória humanizada, um processo de memória humanizada seria mais ou menos isso aqui, esse daqui seria o objetivo final que a gente iria querer alcançar com a memória de um agente de IA, isso daqui é um print screen do aplicativo Obsidian, a proposta desse aplicativo é basicamente, demonstrar a memória humana conectando os pontos das nossas conversas, não das nossas conversas exatamente, é mais um aplicativo de lembretes, então vamos falar, ah, eu tive uma ideia, hoje eu tive uma ideia, sei lá, ah, eu quero explorar as cataratas do Niágara, e eu coloco lá no Obsidian como se fosse um lembrete mesmo do meu telefone, e daí eu começo a falar sobre outras coisas, a catarata do Niágara, e vai conectando uma coisa com a outra, e não precisa ser só sobre a catarata dos Niágaras, só que todas as informações que você vai jogando nesse banco aqui do Obsidian, vai se conectando uma hora ou outra, se as informações se conectarem, isso daqui é a melhor maneira da gente enxergar a nossa memória, se conectando uma com a outra, é o melhor que a gente consegue enxergar hoje, em projetos de IA, de agentes de IA, a gente não consegue ter essa percepção, porque a nossa interação com o agente não é visual, ela é ou por texto ou por voz, dependendo do agente de IA que você está criando, eu já mostrei aqui no meu canal do YouTube alguns agentes de IA por ligação, que você consegue fazer vendas através de ligação, mas também tem os projetos de IA que fazem vendas no WhatsApp, mas para IA conectar todos esses pontos que foram mencionados na conversa, como se fosse uma teia de aranha, a gente não tem percepção visual disso daqui acontecendo na memória do nosso agente de IA, e nem ela foi projetada para acontecer assim, a memória dos agentes de IA foi projetada para ser mais ou menos assim, um banco vetorial, você pega aqui uma informação, pega outra informação aqui, e daí através de semânticas elas se conectam, e aqui pareceu até um bonequinho, eu vou fazer um oi aqui, eu acho que uma demonstração no Google pode ser melhor, deixa eu colocar aqui, vector database, vamos lá, aqui, era esse exemplo aqui que eu estava procurando para vocês, os pontos se conectam através desse gráfico linear, mas a memória do ser humano não é nada linear, então se a gente quer replicar esse tipo de memória seria mais ou menos algo como esse daqui, que a gente consegue encontrar nessa ferramenta chamada Zap AI, olha só, isso daqui é um exemplo da memória de um agente de IA, que eu troquei duas mensagens apenas, e essa teia de aranha aqui já vai tomando corpo, vai tomando vida, vou mostrar aqui a conversa que eu tive com ela, foi mais ou menos assim, meu filme favorito é Lobo de Wall Street, e ela começou a falar, Lobo de Wall Street é um filme incrível, dirigido por Martin Scorsese, estrelado por Leonardo DiCaprio, como Jordan Belfort, eu falei, me indique filmes nesse parâmetro, Lobo de Wall Street, O Homem que Mudou o Jogo, Mono e Ball, ok, já ouvi falar, A Rede Social, muito bom, é top esse daí, eu assisti duas vezes já, Os Estagiários eu não assisti, Wall Street eu não assisti, A Grande Aposta eu assisti, é legal, e o Lobo de Wall Street de novo, mas bem, só com esse tipo de conversa aqui, a gente já conseguiu criar uma teia de aranha, com as recomendações dos filmes que a IA passou pra gente, e do que eu conversei com a IA, eu conversei sobre o Lobo de Wall Street, ela deu isso aqui como ponto de partida, ela fez algumas conexões, e trouxe algumas sugestões, então eu falei, meu filme favorito, e tá aqui, Favorite Movie Of, e aqui o nome da minha memória, esse daqui é o nome do meu usuário dentro dos apps, cada usuário vai ter uma memória dessa daqui, e ela tá se conectando comigo, é uma abordagem muito mais centrada na pessoa, no usuário em si, ela vai conectar os pontos na hora da conversa. Então, bem, esse daqui é o tipo de resultado final que a gente quer alcançar nos nossos projetos de IA, e quem se beneficia disso, né, quais são os casos de uso, quais são as estratégias que a gente pode utilizar, quer dizer, quais são os exemplos que a gente pode ter com essa estratégia.

De fato, a indústria que mais se beneficia com isso daqui é a indústria médica, médicos se dão muito bem com isso daqui, você tem uma memória ampla do seu paciente, colocar um IA pra trabalhar nesse paciente, seja na parte de triagem, ou consultas, agendamento, agendamento de consultas. Em ambos os casos, a gente já trabalhou aqui dentro da Marvin AI, a gente já trabalhou na automatização de agendamentos pra clínicas médicas, e a gente tá trabalhando num aplicativo, eu não vou entrar muito a fundo nisso, mas a gente tá criando um aplicativo pra telemedicina, onde a gente vai colocar um IA responsável pelo processo de triagem, e essa memória personalizada aqui acabou fazendo toda a diferença no resultado final que a gente vai entregar, porque quando a gente cria um aplicativo, a gente fica muito preocupado com o ferramental das coisas, será que isso vai funcionar, será que aquilo não vai funcionar, será que o aplicativo vai dar bug, mas honestamente, com essas ferramentas no code, a estrada fica muito aberta pra gente, e ter garantia de entregáveis fica mais fácil, por isso que a gente deve ir mais longe no resultado final, na proposta de valor do aplicativo, então a gente podendo ir mais a fundo na proposta do aplicativo, a gente vai nessa triagem personalizada. São dois casos de uso que se beneficiam bastante.

Outro caso de uso que se beneficia bastante são prestadores de serviço, em geral, tá? Prestadores de serviço em geral, por exemplo, vamos colocar aqui do primeiro caso uma manicure, essa manicure, não uma manicure, um salão de beleza, vamos colocar aqui um salão de beleza. Salão de beleza você fez, sei lá, você colocou mega-ré e você tem que fazer manutenção, então na hora que você falar com o agente de IA sobre ah, eu quero fazer a manutenção do meu mega-ré, ah sim, você fez a sua última manutenção há quanto tempo atrás, já tá na hora de você fazer a manutenção, em cima disso você já comprar um produto, porque na fase que você tá hoje, você precisa de um tratamento mais especializado naquilo, nisso, naquilo outro, então prestadores de serviço para vender o serviço deles, geralmente eles têm um acompanhamento contínuo, por exemplo, uma assessoria de marketing, presta assessoria contínua para os clientes deles, então são empresas que estão acompanhando o cliente ali há seis meses no projeto deles e falar tão a fundo assim sobre projeto vai deixando a IA cada vez mais rica, então se você, uso interno, uso interno, administração, né, eu pulei aqui a minha coesão narrativa, mas para usos de administrações, uso administrativos, como gestão de projetos, eu vou colocar aqui gestão de projetos, funciona muito bem, por exemplo, você tem uma IA que tá acompanhando o projeto interno da sua empresa, cada ideia que você troca mais com essa IA vai ficando mais valiosa, ainda mais se ela tá integrada num aplicativo de gestão de projetos, tipo o ClickUp, fica ainda mais rico a experiência para o administrador daquela empresa, esse daqui é o típico projeto que você vai vender para CEOs, CFOs, COOs, o pessoal mais C-level e geralmente de empresas mais qualificadas, esse daqui você tem uma barreira de entrada menor, não esse, esse daqui para criar um aplicativo, não é todo mundo que cria um aplicativo, né, geralmente esse tipo de aplicativo tem um investimento considerável para você construir eles, mas clínicas médicas zilhões no mercado se beneficiariam muito desse tipo de projeto aqui, de IA com memória longa, prestador de serviços zibilhões, tanto no mercado B2B quanto no mercado B2C, eu falei das agências de marketing, eu falei do salão de beleza, você pode ir mais a fundo ainda, por exemplo, escritório de advocacia se beneficiariam bastante, gestão de projetos, isso daqui é incrível para gestão de projetos, você tem um mapeamento, você tem um conselheiro, você criar um agente de IA que é um conselheiro na gestão de seus projetos e em cima disso ele está linkado no seu click up para ver o progresso do seu projeto, porque a maioria do que os CEOs tem hoje, é por exemplo, ele tem um GPT, ah, eu utilizo inteligência artificial na minha empresa, o cara tem um GPT que ele conversa para tomar a decisão do negócio dele, ok, ele já deu um passo à frente, ele está na frente de muita gente que não está utilizando inteligência artificial, que é loucura hoje, mas por outro lado, você também não está indo muito longe, meu amigo, essa IA ela só está te dando conselhos baseado no que ela tem de informação da internet e já é muita coisa, você já está tendo acesso a muita coisa, mas você poderia estar extraindo muito mais potencial de uma IA treinada especificamente para o seu negócio, esse é que a gente está especializado em criar hoje, integrar as suas ferramentas, integrar o teu banco de dados, integrar o teu enriquecimento de informações para criar um projeto personalizado para ti, então isso daí é a nossa especialidade hoje de projeto. Enfim, vamos lá para a parte das trincheiras, como a gente cria esse tipo de projeto utilizando essa ferramenta do Zap AI.

Bem, a integração aqui dentro do N8n ela é muito intuitiva, felizmente, então essa daqui é a parte mais curta do vídeo, porque a gente literalmente tem um Node nativo aqui dentro do N8n, a gente tem um módulo nativo, se eu venho aqui em Memory, tem o Zap e eu vou criar uma credencial nova aqui, a gente só precisa colocar a chave API e Cloud, colocar aqui como Cloud. Eu vou pegar aqui a chave API do meu Zap, eu venho aqui em Project, primeiro você tem que criar uma conta no Zap, isso daí é obrigatório, para criar uma conta é gratuito também, mas você vai ter um projeto, então colocando aqui, se você é desenvolvedor de IaaS ou engenheiro de IaaS, seria uma conta para cada cliente e você vai conectar essa conta no N8n criando uma chave API. Eu vou criar uma chave API aqui em Add Key, colocar Demo YouTube Zap, Create, copiei, ok, salvei.

Eu coloquei Cloud, isso daí já te leva a pensar que tem um plano self-hosting, open source, local e sim, tem. A minha abordagem sobre ferramentas self-hosting, eu utilizo o N8n self-hosting, mas a minha abordagem para ferramentas self-hosting costuma ser limitar ao máximo o uso de self-hosting, eu só uso em casos muito específicos, como é o caso do N8n, porque esses casos aqui de bancos de dados, muitas pessoas têm uma abordagem mais escassa da coisa, de falar, ah, eu vou deixar tudo no self-hosting porque daí eu não vou ter que pagar nunca e todos os dados vão ser meus. Bem, é bem legal que você tenha esse senso de propriedade das suas coisas e que você não queira pagar ou investir, mas pensando por outro lado, dados como histórico de conversa costumam escalar muito rápido, conversas costumam escalar, esses projetos de ar foram feitos para as pessoas conversarem, então logo logo você vai ter milhões de dados e vai estar tudo no seu colo, vai ser tudo responsabilidade sua, não seria mais valioso você delegar essa responsabilidade para uma ferramenta especializada nisso, por exemplo, 5 dólares, só quando você precisar, porque a maioria das pessoas pensam que, ah, eu vou colocar no self-hosting porque eu nunca vou ter que pagar, sendo que ferramentas tipo o SupaBase tem um plano gratuito muito amplo, vai demorar muito para você ter que pagar, eu vou falar por mim mesmo, nunca tive que pagar o SupaBase, até em projetos grandes que eu entrei a gente consegue utilizar o SupaBase gratuito.

Então, bem, a maioria das pessoas tem essa abordagem escassa, eu não vou muito longe nisso daí, eu não alimento muito isso, eu tenho alguns vídeos sobre self-hosting do N8n aqui no meu canal do YouTube, porque realmente é uma ferramenta que o plano cloud não vale a pena, mas em ferramentas que o plano cloud é muito abrangente, tipo o Zap, SupaBase, vale a pena utilizar, então eu estou utilizando, só ver aqui a minha credencial, esse daqui é o framework de agentes de IA com a memória longa que eu mostrei para vocês, pessoal, e bem, eu fiz aqui um teste da conversa comigo mesmo, pessoalmente falando, eu vou mostrar aqui aonde que eu consigo ver aquele gráfico, eu venho aqui em usuários, eu vou pegar a última conversa que foi trocada entre eu e a IA, cada usuário tem um ID, esse ID pode ser um número de telefone, por exemplo, se você está colocando seu agente de IA no WhatsApp, eu vou clicar só no ID, vou ver o gráfico, e olha só, que bonito, e tipo assim, cada vez mais a gente vai trocando informação, isso daqui vai se alimentando e os pontos vão se conectando no futuro, mas bem, esse caso ultrapessoal que eu dei para vocês foi por propósito demonstracional do vídeo, para vocês verem a ferramenta, como funciona a maneira de pensar, como é a integração, mas eu focaria totalmente nos casos de uso, para entregar valor nos projetos de IA, por exemplo, uma IA de triagem, que faz um processo de triagem, como é que eu posso explicar o que é um processo de triagem? Aquelas perguntas que você faz quando você está no hospital, que o enfermeiro faz para você quando você chega no hospital, então aqui, ao invés de fazer uma série de perguntas roteirizadas, a IA faz uma série de perguntas mais personalizadas, tipo, se é um paciente que tem histórico de ansiedade, você faz umas perguntas mais ou menos assim, ah, você comeu bem hoje, você acordou com dor no estômago, será que essa dor no estômago não tem a ver com a sua crise de ansiedade que você costuma ter de madrugada, esse tipo de coisa, consultas, eu nem preciso falar, tipo, ah, você quer reagendar uma consulta para ver aquela coisa lá do seu olho que estava coçando, você quer ver sobre a sua crise alérgica que você teve da última vez, esse tipo de coisa. Prestador de serviço, muito mais para impulsionar uma venda, aquelas coisas de remarketing que eu falei para vocês, você pode até criar automações de remarketing, lembretes para forçar vendas, para tentar extrair mais vendas, então isso aqui vai muito bem com automações também, são várias as automações, as possibilidades são várias, realmente, a melhor sugestão que eu tenho para quem quer implementar esse tipo de negócio aqui para a empresa é prestar uma consultoria de automação ou IA, prestar uma consultoria de automação e IA com um profissional, então eu me deixo a disponibilidade para esse serviço também, eu deixo a minha agenda na descrição do vídeo para você ter uma consultoria de IA e entender como ter mais resultado na sua empresa, implementando automação com IA, e é isso daí pessoal, esse daqui foi o aprendizado que eu quis passar para vocês hoje, esse framework de Memoryshed, eu agradeço demais por você ter vindo até aqui e ter investido o seu tempo que é a coisa mais valiosa que você tem, então por isso eu sou muito agradecido, tchau tchau, um beijo, até mais.

**Transcrito por** [**TurboScribe.ai**](https://turboscribe.ai/pt/?ref=docx_export_upsell)**.** [**Atualize para Ilimitado**](https://turboscribe.ai/pt/subscribed?ref=docx_export_upsell) **para remover esta mensagem.**