**6 Truques para o seu Agente IA parar de Alucinar no N8n!**

**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? Eu, Pires, aqui, nesse vídeo, vou apresentar pra vocês 6 dicas rápidas pro seu agente de IA parar de alucinar aqui dentro do N8n. Se você tá criando agentes de IA como chatbots e colocar pra automatizar vendas, suporte ao cliente ou qualquer outra atividade interna pras empresas, você pode passar um pouco de dor de cabeça pra fazer esse tipo de projeto funcionar se você for um pouco mais amador, se você tiver poucos meses nesse mercado. Eu tô há 18 meses criando esses tipos de agentes de IA pra outras empresas, eu tenho uma agência de IA chamada Mavren AI Solutions, e eu documento alguns dos meus cases de sucesso aqui no YouTube já faz um ano e meio, e eu vou mostrar pra vocês essas 6 dicas valiosas aqui dentro do N8n, beleza? A primeira dica que eu queria mostrar pra vocês é sobre a escolha do modelo de LLM, eu vejo muita gente batendo cabeça nisso daí, por mais que seja algo direto ao ponto, eu já mostrei aqui que eu tô utilizando o chat GPT nesse caso específico, mas tem diferentes casos de LLM que eu escolho.

Bem, aqui no chatmodel, que é o modelo de LLM que a gente vai escolher pro nosso agente de IA, você tem uma gama de opções, nem sempre foi assim, nem sempre teve tantos modelos, a gente tá na melhor fase das LLMs, eu poderia dizer. O Google Gemini tá fazendo lançamento atrás de lançamento, e cada lançamento é uma pedrada, os caras tão indo muito bem, o chat GPT tá sendo pressionado a lançar novos modelos, e os caras tão mandando bem também, não tão surpreendente quanto o Google, isso porque o Google saiu do 0 pra 1, saiu do 0 pra alguma coisa, o chat GPT já tá sendo sólido há um bom tempo, mas a gente viu alguns meses atrás a descoberta do DeepSeek, o Cloud sempre faz lançamentos bombásticos, e eu vou falar pra vocês qual que eu utilizo nos meus modelos de IA, dependendo do projeto e porque eu escolho. Bem, Cloud AI é um modelo excelente pra automações de vendas, copywriting, marketing, se você quer escrever uma copy ou você quer fazer um post pras suas redes sociais, o Cloud é o melhor modelo de LLM pra escrita, eu venho falando isso há um bom tempo no meu canal do YouTube, já faz mais de um ano que eu venho batendo nessa tecla, inclusive eu tenho um modelo valiosíssimo de automação aqui no meu canal, que é o meu modelo de propostas automatizadas, eu basicamente envio propostas automáticas pros potenciais clientes da Maven AI em pouquíssimos minutos, e quem escreve essas propostas é o Cloud AI, isso porque ele tem um modelo de escrita bem melhor, eu posso compartilhar aqui esse modelo de proposta rapidamente pra vocês, deixa só eu logar no Make, eu utilizo o Make pra fazer essa automação de proposta porque não é exatamente um agente de IA, como eu falei pra vocês, é uma automação de proposta, tem uma mudança de mentalidade entre um e outro, olha só, veja bem, eu tenho aqui no início o chat GPT e do resto pra frente Cloud AI, isso porque eu tenho uma mentalidade mais definida pra quando utilizar cada LLM, esse daqui é a LLM do chat GPT, eu utilizo pra pegar o número de telefone do potencial cliente e identificar o país dele através do número de telefone, isso porque quando a gente vai disparar uma mensagem no WhatsApp a gente precisa do DDI e do DDD, tudo bem certinho, eu não posso cometer um erro nisso senão a automação vai falhar, e por isso eu coloco o chat GPT pra identificar esse país e o resto da proposta quem escreve é o Cloud AI, por exemplo, o parágrafo, a página de visão geral do projeto Cloud AI, uma página sobre resultados estimados Cloud AI, eu tenho um vídeo cobrindo só essa automação aqui, ela me traz um retorno imediato, ela me trouxe um retorno imediato de mais 10 mil reais por mês assim que eu implementei ela, não é um vídeo só sobre ela, esse daqui, mas eu tenho um vídeo só sobre ela no meu canal do YouTube, eu vou deixar o card, tipo assim, eu tava numa época faturando 5 mil por mês e assim que eu implementei essa solução meio que eu tripliquei o meu faturamento por conta de como você aquece o seu processo de vendas e é isso que o cliente tá esperando de você, e para pra pensar, se você é uma empresa de automação, você tem que ter os seus processos automatizados, ok, então eu uso o Cloud AI pra escrita e vendas, agora se eu tiver utilizando um chatbot de vendas, eu vou utilizar o Cloud AI, a resposta é não, eu utilizo o chat GPT sempre pra qualquer chatbot, seja um chatbot de vendas, seja um chatbot pra operações internas de uma empresa, chatbot pra suporte ao cliente, isso porque o chat GPT é um modelo ultra sólido, ele consegue aguentar um volume de requisições muito tranquilo e o custo da API é muito menor do que o do Cloud AI, o Cloud AI tem um custo de requisição maior e não aguenta tantas requisições assim, o servidor costuma cair quando é pra fazer uma chamada à API e o chat GPT aguenta toda pancada, na verdade eles tão indo pra esse caminho de abrir o mercado pra ser uma API, pra ser a melhor API do mundo e os caras aguentam, tão jogando esse jogo, tão monetizando e tão monetizando em volume, porque cada vez mais eles diminuem os preços.

Agora, Google, a Google tá jogando o mesmo jogo do chat GPT, inclusive eles são vastamente conhecidos por jogar esse jogo ao longo das décadas e eles sempre ganham, é difícil você ver a Google perdendo quando se fala de volume, só o Google Search, a pesquisa do Google recebe 2 bilhões de visitas na página por dia, então isso se não for em uma hora ou em poucas horas, mas os caras estão acostumados a jogar o jogo do volume, eles definitivamente estão olhando pra isso, eles estão sempre deixando os modelos do Google Gemini abertos pra testes e a gente sempre consegue testar e diga-se de passagem tá performando muito bem. Na criação de chatbots, o Google Gemini foi o melhor que eu encontrei, ele conseguiu ter as melhores respostas, o agente já ficou mais humanizado, ponto, parágrafo, ainda não tá aguentando todo o volume de requisições, tá bugando, então a gente vai ter que esperar alguns meses pra poder colocar essa belezinha aqui em produção e ter os nossos chatbots rodando no Google, que agora tá sendo o melhor modelo pra chatbots, mas por conta de não ter essa solidez de aguentar volume, a gente recorre ao chat GPT, que também não tá mal, então esse é o modelo que eu utilizo, pra ser mais específico com vocês, eu utilizo o GPT 4.1 Mini, deixa eu encontrar ele aqui, eu utilizo ele no meu chatbots, isso porque é o modelo mais recente que a OpenAI lançou, eles lançaram o GPT 4.1, 4.1 Mini e 4.1 Nano, o 4.1 é o meio termo deles, é pra tarefas mais rápidas e mais pensativas, então se você tem um prompt que a sua IA precisa pensar um pouco mais, o 4.1 vai funcionar muito bem, o 4.1 Mini vai funcionar muito bem, o 4.1 definitivo é pra quando você realmente precisa pensar muito, você dá instruções muito longas pra sua IA, só que tem um custo maior do que o 4.1 Mini, então a gente costuma deixar o prompt mais enxuto e utilizar o 4.1 Mini, essa é a realidade nossa aqui, na Maven AI Solutions, a minha agência de IA, e a gente tá indo bem com esse modelo aqui de LLM. A próxima dica que eu quero passar pra vocês é extrair o máximo possível do seu modelo de LLM, se você tá utilizando o chat GPT ou o Google Gemini, beleza, é uma LLM boa, mas você pode extrair mais recursos dela, você pode deixar ela mais alinhada com o que você quer, tipo, a maioria das pessoas vê diferença do resultado de uma pra outra e fala, ah, o GPT tá mais humanizado, o Gemini tá mais humanizado, então o GPT não costuma me obedecer, o Gemini alucina, na verdade, é como você monta toda a estrutura que vai definir o seu resultado, não é só a LLM, a LLM é um dos componentes que vai definir o seu resultado.

O próximo passo que você tem que fazer é, pelo menos, dar alguns parâmetros pra tua LLM, por exemplo, se você quer que ela siga mais o teu prompt, siga mais restritamente o teu prompt, uma boa dica, você vem aqui em options, abrindo aqui, options, coloca sampling temperature, pra você definir a temperatura do seu agente IA, essa temperatura vai de 0.1 até 1.0, isso se 0.1 é o mais estrito possível, o mais restrito possível, e o 1.0 é o mais livre possível, então se você deixar 0.4, você tá mais perto de deixar ele restrito, você tá fazendo ele seguir mais o prompt, e se você deixar ele no 0.9, ele tá mais livre, por padrão ele vem no 0.7, eu gosto de colocar no 0.9, porque eu gosto de deixar a IA mais generativa, mais livre, mais humanizada, porque desde o início, o princípio desses projetos aqui de IA, de LLM, é ser uma conversa natural, então a gente consegue, pelo menos na minha visão, extrair mais resultados fazendo o que ela foi feita pra fazer. Você pode extrair mais coisas também aqui dentro das opções da sua LLM, como por exemplo esse maximum number of tokens, você escolher o limite máximo de tokens que o seu agente de IA vai utilizar, de 0 a 200, isso pra você ter um controle maior dos seus gastos no fim do mês da API. Bem, eu deixo no padrão mesmo, mas você pode definir como você quiser pra controlar os seus gastos, isso daqui não é 80-20, isso não vai definir se a sua IA alucina ou não, mas eu vi que era uma dica valiosa pra deixar aqui nesse vídeo.

O próximo passo que é muito importante pra você deixar a sua IA humanizada, sem alucinar nas conversas, é você decidir uma boa memória pro seu agente de IA. E quando eu falo memória, não quer dizer o banco de arquivos que você vai colocar dentro da IA, mas sim o framework de memória que vai fazer ela lembrar das conversas que você teve com ela, isso daí é importante. E aqui a gente tem boas opções até dentro do N8n, eu vou ser bem honesto com vocês, de todas que eu testei, eu gostei de todas.

No início eu utilizava o Redis, ChatMemory, isso porque os meus agentes de IA, essa automação padrão aqui de IA, ela já usa o Redis pra fazer um buffer de mensagens, se você não tá habituado a esse termo, buffer de mensagens é um framework que a gente faz pro nosso agente de IA ficar humanizado no WhatsApp. Como que humanizado? Ele junta uma fila de mensagens, então se o usuário enviou uma mensagem nos últimos 8 segundos, ou 3 mensagens nos últimos 8 segundos, a gente junta todas aquelas mensagens e envia de uma vez só pra IA responder, pra ela não ficar respondendo uma por uma e não ficar aquela coisa robotizada, igual é o ChatGPT, a gente pode enviar mais mensagens de uma vez só, mas tem esse delayzinho que você escolhe o delay pra juntar as mensagens. Por padrão eu deixo 8 segundos, você pode deixar um padrão menor, depende do público também que você tá escolhendo.

Por exemplo, a gente tem um background aqui na nossa agência de IA que a gente trabalha com escritório de advocacia, é uma parte significativa da nossa carteira de clientes, da nossa carteira de parceiros. E esse pessoal atinge um público, por exemplo, de direito previdenciário, é um público com uma faixa etária maior, então eles demoram mais tempo pra enviar mensagens e o delay tem que ser maior, então a gente coloca 10 segundos de delay, esse tipo de coisa. Mas voltando aqui pro fundamento, pros nossos agentes de IA, a gente testou o Redis, a gente testou o Postgres Chat Memory, esse agente de IA aqui tá utilizando o Postgres Chat Memory, é um agente de IA que eu tô deixando de uso alternativo há muito tempo, mas recentemente eu venho testando o Zap Memory e pra mim ele foi o melhor memória de agente, isso porque ele segue um framework de memória humanizada, ele guarda as conversas do usuário na mente do agente IA como um humano organiza as memórias dele.

Eu tenho um vídeo só sobre isso, eu vou deixar aqui no card pra você entender mais a fundo sobre o que eu tô falando e eu também ensino a instalar, não instalar, a integrar o Zap Memory dentro do seu N8n. O próximo passo pra você deixar o seu agente de IA sem alucinar é você escolher bem as tools do seu agente de IA. As tools é o que dá a autonomia pro nosso agente de IA fazer ações.

Então, por exemplo, você quer um agente de IA que cria propostas pra você, ele tem que ter uma tool conectada ao seu aplicativo de propostas, no meu caso é o PandaDoc, mas o PandaDoc não tá aqui dentro do N8n, por isso que eu faço as minhas propostas com o Make.com, se isso é algo que vai tirar a dúvida de vocês antecipadamente ou não. Mas bem, você pode ver que tem vários aplicativos dentro do N8n, por exemplo o ClickUp, o ClickUp é significativo aqui na nossa agência, a gente usa o ClickUp pra muita coisa e você deixa o seu agente IA conectado ao ClickUp se você quer fazer automações dentro do ClickUp. Qual o problema que eu vejo as pessoas tendo com agentes de IA alucinando? Eles não definem muito bem os parâmetros dentro da tool.

O que é que eu quero falar por isso? Eu vou trazer aqui o exemplo do Google Calendar, que é o que eu mais vejo o pessoal alucinando. Por exemplo, eu quero criar um agendamento, eu tenho que colocar a hora de início desse agendamento e a hora de fim e o pessoal sofre pra instruir o agente de IA nessa hora de início e nessa hora de fim, até porque por padrão já vem um código aqui que pode ficar um pouco esquisito pra você. Mas bem, a melhor maneira que eu desvio esse obstáculo aqui de código pro pessoal que é iniciante é utilizando uma AI.

O N8n tem esse modelo de AI pra te ajudar, pra você não ter que ver código o máximo possível. Então, se você clicar aqui nesse botão ao lado e nesse botão ao lado, você vai deixar que a IA decida o horário de início e o horário de fim. Mas você tem que dar um prompt pra essa IA pra ela não fazer besteira.

Então, por exemplo, o horário de início seria o horário que o usuário pediu pro agente de IA e daí eu coloco isso aqui no prompt. Horário que o usuário escolheu com o agente de IA. E aqui o horário de fim eu coloco duas horas após o horário de início que o usuário escolheu com o agente de IA.

E daí eu desvio daquela dor que eu teria de programar isso daqui dentro do N8n. Eu teria que lidar isso da seguinte forma. Aqui você já vê que tem um código.

Eu teria que escrever tudo isso daqui e essa estrutura de código pode ser um pouco difícil se você tá iniciando dentro do N8n. Eu tive essa curva de aprendizado, me levou um pouco de tempo pra criar agentes de IA melhor, com tools funcionando, mas isso daqui ajudou bastante. Eu já não tenho mais essa dor por conta dessa ferramenta do N8n.

Outra coisa que eu vejo o pessoal errando na hora de fazer o seu agente de IA ter tarefas autônomas é que eles assumem que um tool só vai resolver o problema deles. Então, por exemplo, se você quer atualizar o seu cliente lá no ClickUp, você quer atualizar o número de telefone dele e você quer enviar um e-mail pra ele, as pessoas assumem que só a tool do ClickUp vai fazer isso. Idealmente você teria que fazer uma tool do ClickUp e uma tool do Gmail.

Então, ficaria mais ou menos assim o seu agente de IA. Calma aí. Assim ficaria a estrutura do seu agente de IA se você quisesse que ele se conectasse ao ClickUp e ao seu Gmail.

E pra fazer ele ativar isso em sequência, você teria que deixar o prompt muito bem definido, passo a passo, igual esse prompt aqui, que tá seguindo um passo a passo, mas isso não iria deixar ele imune de pular um passo. Então, bem, o que eu recomendo que você faça é, se você precisa fazer mais do que uma automação, por exemplo, nesse caso de atualizar o ClickUp e enviar um e-mail de uma vez só, eu recomendo que você utilize a tool call n8n workflow tool, porque você cria um workflow de automação e você consegue criar nesse fluxo de automação tudo em uma vez só. Como que eu vou explicar isso daí pra vocês? Aqui, ó. Então, ficaria assim o resultado final que eu quero atingir.

Eu iria falar pro meu agente de IA ativar a tool, eu daria o nome pra essa tool, por exemplo, Lead Qualificado. E quando o agente de IA ativasse a tool Lead Qualificado, essa tool iria ativar esse workflow e esse workflow iria atualizar o nosso cliente lá no ClickUp e enviar um e-mail. Isso aqui eu coloquei só de demonstração pra vocês entenderem que ele iria seguir uma sequência lógica e não iria depender tudo do prompt.

Um erro que eu vejo as pessoas cometendo é deixar tudo para o prompt resolver, sendo que a maioria das vezes está na estrutura de automação que você está montando em si. Ok, outra dica valiosa pra vocês, isso aqui é uma dica que salvou o meu projeto de IA SDR, é você definir os parâmetros dentro do seu agente de IA. Eu não falo dentro da sua LLM, na sua LLM aqui já está definido, mas dentro do seu agente de IA, assim que você colocou o prompt, tem algumas coisas que você pode adicionar pra ele seguir mais a risca o seu prompt, além da LLM em si.

Eu digo essa max iterations aqui. Uma dor que eu estava tendo no meu projeto de IA SDR, esse daqui, é que quando o lead estava qualificado, ele ativava várias vezes essa tool de lead qualificado. E essa tool de lead qualificado, se eu encontrar aqui pra vocês, deixa eu procurar ela.

Essa é a tool de lead qualificado, ela tem várias ações dentro do meu workflow, ela envia uma mensagem de follow-up no WhatsApp, ela atualiza minha CRM no Superbase, ela envia outra mensagem de WhatsApp através de um outro número de telefone. Então, isso daqui são várias coisas pra fazer de uma vez. Eu utilizei o Call Another Workflow Tool, aquela sugestão que eu dei pra vocês, eu coloco ela em prática, por isso que eu recomendo, mas, bem, ela estava ativando essa tool aqui mais do que uma vez.

Quando o lead estava qualificado, o lead falava, ok, pode me passar pra um especialista, e ela ficava ativando essa tool repetidas vezes e ficava flodando esse usuário de mensagens. Ficava algo muito alucinógeno, ela alucinava. Então, pra gente limitar o nosso agente de IA pra ativar a tool só uma vez, a gente vem aqui em Max Iterations, coloca 1 e voilá.

Se eu tivesse conhecido isso há meses atrás, eu não perderia tantos leads igual eu perdi. Ok, e por último, mas não menos importante, eu quero deixar vocês aqui com chave de ouro, fechar com chave de ouro esse vídeo. A principal dica é engenharia de prompt.

Um prompt bem formatado, com informações claras e linguagem natural pro seu agente de IA vai te trazer resultados melhores. Vamos lá, mais uma vez, eu recebo algumas consultas no meu direct, ou então dentro da minha comunidade, ou então em qualquer canal de contato, as pessoas me perguntam qual é o melhor formato de prompt. Eles se apegam que ou o formato de markdown vai salvar o prompt deles, ou o formato de XML vai salvar o prompt deles, mas não é o formato que vai salvar o seu prompt.

Eu utilizo o formato de markdown há um ano e meio, e já surgiram vários formatos novos que alguns criadores de conteúdo trouxeram aqui no YouTube, mostrando que era a salvação do seu agente de IA, mas a salvação não está no formato, está no prompt em si, isso é algo bem lógico, se você parar pra pensar um pouquinho só fora da caixa, você está buscando por salvação, você está buscando por um objeto precioso. Então, bem, eu gosto do markdown porque é mais limpo pra eu corrigir se tiver algum erro, o prompt em XML é difícil de eu ler, é difícil de ajustar se tiver algum erro, já o markdown eu consigo apagar e reescrever quantas vezes eu quiser. É bem fácil, na verdade, isso aqui é bem natural pra gente ver algo em markdown.

Markdown é essas hashtags aqui que ficam, tipo, hashtag role seria função, hashtag task seria a tarefa do agente IA, hashtag specifics seria as especificidades do projeto, hashtag contexto e as notas aqui. Eu sigo esse padrão de prompt, normalmente eu coloco as notas no final, por conta de algo chamado Lost in Middle Effect, isso daí já é algo documentado entre cientistas de IA, que se você coloca as informações no início ou no fim do agente IA, essas são as informações que ele vai reter ao máximo, e as informações que ficam no meio podem ficar mais perdidas, tem mais chance de ficar perdida. Então, no final, eu coloco as notas, que são a minha última chance de fazer o meu projeto dar certo, então eu coloco várias coisas negativas aqui pra ela não fazer ou não deixar de fazer, como por exemplo, não prometa materiais de apoio, não prometa indicações, só continue com clientes qualificados, não invente nomes, não solicite número de telefone, não mencione tool lead qualificado, só faça uma pergunta por vez.

E bem, eu me esforço muito em deixar a conversa no prompt muito natural, isso porque eu não faço aquele tipo de prompt como se eu estivesse falando com um funcionário, sabe? Não com um funcionário, eu falaria pior do que um funcionário. Quando você reduz a pessoa ao mínimo e você fala, faça isso, faça aquilo, faça aquilo, eu não passo as ordens assim para o meu agente de IA, eu passo as ordens para ele como se eu estivesse conversando. Imagina que eu tenho uma pessoa do meu time da Mava.ni.ai, por exemplo, se eu quero instruir um dos meus vendedores, eu o instruo através de uma conversa e não falando, faça x, faça y. Por mais que você sofra para seu objetivo, por exemplo, alguns textos ficam mais longos, mas fica melhor para a IA entender, porque mais uma vez, esses modelos de LLM foram feitos para simular uma conversa real, então o seu prompt tem que ser uma conversa real com o seu agente de IA.

E é assim que eu deixo o meu prompt, beleza? Essas aqui foram as dicas valiosas que eu queria deixar para vocês, testem aí na casa de vocês, testem nos seus projetos de IA e se derem resultados, comentem aqui, por favor, para eu ficar sabendo se esse vídeo ajudou você e se alguma dica já destravou a tua cabeça, já deixa o like e se inscreve no canal para ficar ligado em mais dicas que eu posto aqui, beleza? Tamo junto, pessoal, muito obrigado pelo seu tempo, que é a coisa mais valiosa que você tem. Tchau, tchau!

**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.**