

O Diagnóstico Era a Parte Fácil

À medida que as máquinas aprendem a diagnosticar melhor que os médicos, a fronteira da saúde pessoal se revela algo mais antigo e humilde: uma memória fiel.

Rafael M. Ehlers · rafaelmehlers.com.br · Julho de 2026 · doi.org/10.5281/zenodo.21350119

Uma nota ao leitor. Este ensaio se apoia numa única aposta, e prefiro expô-la na primeira linha a enterrá-la: a de que o problema difícil da saúde pessoal deixou de ser o conhecimento médico há algum tempo, e passou a ser a continuidade e a *fidelidade* do registro que uma vida deixa para trás. Se isso estiver errado, se a capacidade ainda for a restrição determinante, o argumento desaba, e é o que deve acontecer. Julgue-o contra o seu próprio último ano de saúde, não contra um benchmark.

A inversão

Durante quase todo o século passado, o gargalo da medicina foi a expertise. O conhecimento vivia em algumas poucas cabeças treinadas e em alguns livros pesados, distribuído de forma desigual e caro de alcançar. Quase tudo o que construímos — hospitais, periódicos, cadeias de encaminhamento — era uma máquina para mover expertise escassa em direção às pessoas que dela precisavam.

Esse gargalo está se fechando, mais rápido do que a cultura absorveu. Em 2025, uma equipe da Microsoft submeteu modelos de linguagem a 304 dos casos diagnósticos mais difíceis que o *New England Journal of Medicine* publica, reformulados de modo que o modelo, como um clínico de verdade, tivesse que fazer perguntas e pedir exames um passo de cada vez, em vez de ler um resumo já arrumado. Um sistema orquestrado alcançou cerca de 80% de acurácia diagnóstica. Os médicos generalistas usados como referência alcançaram cerca de 20%. Quatro para um, e mais barato. O resultado se manteve entre diferentes famílias de modelos; não é um acaso do ajuste de um único laboratório.

Você pode discutir o benchmark, e eu vou fazê-lo daqui a pouco, mas não a direção. Aquilo que passamos um século tratando como escasso e sagrado — o raciocínio diagnóstico — está se tornando abundante. E no momento em que algo escasso se torna abundante, a pergunta interessante é sempre a mesma: qual é o novo gargalo?

O que o benchmark admite em silêncio

A resposta está escondida dentro do próprio estudo que anunciou o diagnóstico sobre-humano.

Repare no que a equipe da Microsoft teve que construir para tornar seu teste realista: um “Gatekeeper” (porteiro), um segundo modelo cuja única função é liberar as informações do paciente, pedaço por pedaço, quando o diagnosticador as pede explicitamente. Eles o construíram porque entenderam que o diagnóstico real não é um problema de raciocínio que recebe um caso completo. É um problema de coleta sequencial de informação. A habilidade que está sendo medida é saber o que perguntar, em que ordem, e quando já se tem o suficiente para se comprometer.

Sente-se com isso. Mesmo dentro do mundo artificial de um benchmark, o diagnosticador é limitado não por quão bem pensa, mas por aquilo que consegue puxar. O raciocínio está resolvido. A recuperação da informação é o jogo.

Agora saia do benchmark. Ali, a informação existe, completa, atrás do Gatekeeper, esperando ser consultada. Numa vida, não. Uma vida não é um caso empacotado. É vivida ao longo de anos, em fragmentos, na maior parte não registrados e semiesquecidos. A frase mais comum em qualquer consultório não é um sintoma. É “acho que começou em julho, ou talvez depois do Carnaval.” A pessoa que mais sabe sobre aquele corpo, a que esteve presente em cada noite ruim, cada refeição e cada dor, é também a menos capaz de contar a sua história.

Então aqui está a inversão por inteiro. Estamos prestes a entregar a todos um diagnosticador que raciocina no nível de um bom médico, ou acima dele. E ele vai ficar sentado, brilhante e pronto, diante de um paciente que não consegue lhe fornecer a única coisa de que ele precisa para ser útil: um relato fiel do que de fato aconteceu, ao longo do tempo.

Certa vez resumi isso num slogan: demos a todos uma biblioteca e esperamos um médico. A frase precisa de uma atualização. O médico chegou. E ainda assim esquecemos de dar a alguém uma memória.

O fio que ninguém segura

Os ricos sempre tiveram uma versão disso. O médico de família que acompanhava uma casa por décadas, que conheceu o avô, o pai e o filho, que conseguia ligar uma queixa de hoje a outra de vinte anos atrás, nunca foi valioso principalmente pelo conhecimento superior. Ele era valioso porque segurava o fio. Ele *era* a continuidade. Quase todo mundo perdeu isso quando a medicina virou uma sequência de encontros de quinze minutos com estranhos, cada um vendo uma fatia, nenhum segurando o todo.

Esta é a vaga em aberto. Não o conhecimento; nesse a gente se afogou. Nem o raciocínio, tampouco, não por muito mais tempo; as máquinas estão tomando isso. O que falta é um registro contínuo e fiel de uma única vida: o substrato sobre o qual qualquer inteligência, humana ou artificial, precisa raciocinar para ser de alguma utilidade para *você* em particular.

Os sistemas que temos não preenchem essa vaga. Um contador de passos coleta e esquece o contexto. Um portal do paciente arquiva documentos que ninguém relê. Um chatbot generalista responde à pergunta de hoje e apaga a lousa antes da próxima. As pessoas já chegam a essas ferramentas exatamente com as perguntas pessoais e longitudinais em que as ferramentas são piores, de forma desproporcional à noite e nos fins de semana, quando o sistema formal está fechado. Mas o chatbot não tem nenhum ontem para ligar ao hoje delas. Cada um deles vive do lado errado do problema. Nenhum segura o fio.

A parte que todo mundo erra: fidelidade

Aqui o argumento fica afiado, e aqui eu acho que a maior parte do que se pensa sobre “IA que lembra” é ingênuo.

É fácil dizer que um agente deveria lembrar de você. É fácil construir algo que armazena. O problema difícil, o que decide se qualquer coisa disso funciona, é se ele lembra *com fidelidade*.

Considere o que construir um registro longitudinal a partir de uma conversa comum de fato exige. Uma pessoa diz, de passagem: “dormi mal, umas cinco horas, acordei com dor de cabeça, provavelmente estresse do trabalho.” Para que essa frase vire memória, um único sistema tem que fazer dois trabalhos opostos ao mesmo tempo: responder como um companheiro acolhedor e, de forma invisível, extrair os fatos registráveis (sono, cinco horas; sintoma, dor de cabeça) e arquivá-los, datados, numa linha do tempo. Depois, semanas mais tarde, algo tem que olhar através de centenas desses fragmentos e fazer emergir o padrão que nenhum dia isolado revela.

Cada passo é uma chance de mentir. Se o extrator anotar oito horas quando a pessoa disse cinco, o erro não permanece pequeno. Ele se acumula. Cada correlação tirada daquela linha do tempo, cada síntese semanal, cada gentil “você tende a se sentir assim depois de...” agora se apoia numa fundação corrompida e herda a corrupção, enquanto *soa* mais confiante, não menos, porque chega vestida de padrão em vez de palpite. Uma memória infiel é pior do que nenhuma memória. Nenhuma memória te deixa onde você começou. Uma memória infiel lava o erro e o transforma em autoridade.

É por isso que eu acho que a verdadeira fronteira da IA para saúde pessoal não é, de modo algum, um problema de capacidade. É um problema de *fidelidade*, e fidelidade é mensurável. Dá para perguntar, concretamente: o evento estruturado corresponde ao que a pessoa realmente disse? As sínteses refletem regularidades reais, ou artefatos que o modelo inventou para parecer perspicaz? Essas são perguntas com respostas, e quase ninguém as está fazendo, porque o campo ainda está intoxicado pela curva de capacidade que, neste domínio, já está achatando rumo à irrelevância.

Quatro maneiras de estar errado

Um argumento que não gera nenhum teste é decoração. Então, de forma falsificável:

Continuidade vence capacidade. Um agente com o seu histórico fiel e um modelo medíocre dá orientação pessoalmente mais útil do que um modelo de fronteira sem nenhuma memória. *Refutado se*, em comparação cega, o modelo de fronteira sem memória empatar com o que conhece você.

O valor está na fidelidade, não no armazenamento. A retenção não auditada vai produzir sínteses confiantes e erradas a uma taxa que corrói a confiança mais rápido do que a memória a constrói. *Refutado se* a acurácia do registro acabar não importando para a confiança ou para os desfechos.

A demanda precede o produto. As pessoas já estão usando assistentes sem memória exatamente para as perguntas pessoais, longitudinais e fora de hora em que esses assistentes são piores. *Refutado se* o uso se mostrar uniforme, impessoal e concentrado no horário de expediente clínico.

O fosso é o registro, não o modelo. O que quer que seja construído aqui será defensável na proporção do histórico fiel que tiver acumulado, não da esperteza do seu raciocínio, porque o raciocínio está virando commodity e o histórico leva anos para existir. *Refutado se* os usuários migrarem sem atrito por mais registro que tenham construído.

A objeção que eu lhe devo

“Você está descrevendo um pesadelo de segurança. Um assistente que acumula todo o histórico de saúde de uma pessoa, identificado e longitudinal, é o objeto de dados mais sensível que se possa imaginar, e o mais abusável.”

Sim. Esse é o custo verdadeiro, e não vou suavizá-lo. A mesma arquitetura que poderia segurar o fio com cuidado poderia segurá-lo como vigilância. Um registro fiel o bastante para ajudar é fiel o bastante para ferir, e dados de saúde identificados e longitudinais são a coisa mais íntima que uma pessoa possui. É por isso que, se algo disso for construído, a propriedade tem que ficar com a pessoa. O fio é dela: portátil, inspecionável, apagável. Não como uma funcionalidade, mas como a precondição da coisa inteira. Quebre essa confiança uma vez e o argumento não perde alguns usuários; ele perde o direito de existir. Levanto a objeção porque ela é a mais forte contra construir isto, e porque ela não toca a tese. Ela eleva o que está em jogo em acertar a tese.

Coda

Comece onde a medicina não para de falhar: alguém numa sala de espera, incapaz de contar a história do próprio corpo. Por cem anos assumimos que o problema dessa pessoa era uma escassez de conhecimento especializado, e construímos catedrais para distribuí-lo. Estamos prestes a concluir o serviço com máquinas que raciocinam como especialistas.

E essa pessoa ainda vai estar ali sentada, incapaz de dizer quando começou.

Porque a coisa que lhe faltava nunca foi o conhecimento do médico, e em breve não será o raciocínio do médico. É a coisa mais humilde de todo o sistema e a última que pensamos em construir: uma memória fiel dela, guardada ao longo do tempo, que amarra a dor de hoje à noite ruim de três semanas atrás e não a obriga a começar do zero toda vez que ela abre a boca.

O diagnóstico era a parte fácil. Ser lembrado, com fidelidade, é a fronteira.

Uma nota sobre o método. Este ensaio afirma uma posição; cada afirmação falsificável acima é para ser levada ao pé da letra, e qualquer uma delas, refutada, o enfraquece. O benchmark de diagnóstico sequencial e os números de aproximadamente 80% contra 20% vêm da primeira referência abaixo; o padrão de uso público, pessoal e fora de hora de chatbots generalistas para saúde vem da segunda.

Referências

Nori, H., et al. "Sequential Diagnosis with Language Models." *arXiv*, 2025.
arxiv.org/abs/2506.22405

Costa-Gomes, B., et al. "Public use of a generalist LLM chatbot for health queries." *Nature Health*, vol. 1, 2026, pp. 689-696. nature.com/articles/s44360-026-00117-x