

MANUAL COMERCIAL E EXPLICATIVO

Tracer-MED

300 perguntas e respostas sobre IA, inferência médica, dados e busca auditável

PERGUNTAS

300 respostas diretas

BLOCOS

**Fundamentos · Dados ·
Mercado · Tracer-MED**

LINGUAGEM

Sem termos técnicos

USO

Comercial e treinamento

1 · Fundamentos: IA, modelo, inferência e RAG

1-10 11-20 21-30 31-40 41-50

2 · Dados, treino, limites e governança

51-60 61-70 71-80 81-90 91-100

3 · O mercado de inferência médica e a análise comercial

101-110 111-120 121-130 131-140 141-150 151-160 161-170 171-180 181-190
191-200

4 · RAG na prática, a stack e a mecânica do Tracer-MED

201-210 211-220 221-230 231-240 241-250 251-260 261-270 271-280 281-290
291-300

1 Fundamentos: IA, modelo, inferência e RAG

O vocabulário mínimo para entender o que o Tracer-MED faz. Não pressupõe conhecimento técnico: cada resposta começa pela ideia em linguagem comum e só depois mostra o efeito prático no hospital.

ANTES DE RESPONDER

Encontrar com prova

A pergunta é comparada com o acervo da instituição e cada documentação que fica de fora recebe uma justificativa calculada.

AO RESPONDER

Injetar o material certo

O modelo de linguagem recebe apenas os prontuários recuperados, com fonte e com a nota de corte registrada.

DEPOIS DE RESPONDER

Conferir e assinar

Cada frase do texto é conferida contra o material enviado, e a consulta entra na trilha encadeada para emissão do relatório.

PERGUNTAS 1 A 50.

1. O que é inteligência artificial?

É um conjunto de programas de computador que fazem tarefas que antes exigiam julgamento humano, como ler um texto, responder a uma pergunta, reconhecer uma imagem ou sugerir um caminho. Ela não pensa como uma pessoa: ela encontra padrões em exemplos e usa esses padrões para produzir uma saída.

2. O que é um modelo de linguagem, como o GPT, o Gemini ou o Claude?

É um programa treinado com uma enorme quantidade de textos para continuar um texto de forma coerente. Quando você faz uma pergunta, ele estima qual é a continuação mais provável e escreve a resposta palavra por palavra. Ele é bom em redigir, resumir e organizar, mas não consulta um banco de dados por conta própria.

3. Por que o modelo de linguagem escreve palavras que parecem certas mas podem estar erradas?

Porque ele foi treinado para produzir sequências de palavras estatisticamente prováveis, não para consultar a verdade. Quando o contexto não traz a informação, ele completa a frase com o que costuma aparecer em textos parecidos. Isso gera respostas fluidas e, de vez em quando, afirmações inventadas.

4. O que é alucinação?

É quando o modelo apresenta como fato algo que não existe ou que não está em nenhuma fonte: um remédio com dose inventada, um exame que não foi feito, uma citação de um prontuário que não existe. O texto parece correto porque está bem escrito. É o principal risco de usar um modelo de linguagem sem conferência.

5. O que é uma 'IA generativa'?

É a IA que produz conteúdo novo a partir do que aprendeu: um texto, um resumo, uma imagem, um código. No caso do Tracer-MED, o foco não é gerar texto bonito, e sim garantir que o material que vai para dentro da geração seja o material certo.

6. O que é inferência?

Inferência é o momento em que o modelo de linguagem é usado para produzir uma resposta. Treinar é ensinar; inferir é usar o aprendido. Na prática, inferência é o que acontece quando você aperta Enviar e a resposta aparece.

7. O que acontece dentro do modelo durante uma inferência?

A pergunta vira números. O modelo processa esses números em várias camadas de cálculo e escolhe, a cada passo, a próxima palavra mais provável diante de tudo o que já foi escrito, até terminar a frase. Cada palavra escolhida entra nas contas da palavra seguinte.

8. O que é um token?

É um pedaço de palavra (uma palavra inteira, uma sílaba ou um sinal). O modelo não trabalha com letras nem com palavras do dicionário: ele trabalha com esses pedaços. O tamanho do texto que cabe no modelo é medido em tokens.

9. O que é a janela de contexto?

É o limite de texto que o modelo consegue ler de uma vez. Tudo o que estiver fora dessa janela é como se não existisse. Se o trecho relevante do prontuário não couber ou não for enviado, o modelo não vai usá-lo, por mais que ele exista no hospital.

10. O que é temperatura?

É um ajuste que deixa a resposta mais previsível ou mais criativa. Temperatura baixa produz respostas parecidas entre si; temperatura alta produz variação. Para uso clínico e auditoria, o comum é usar temperatura baixa para reduzir variação.

11. Por que a mesma pergunta pode gerar respostas diferentes?

Porque o modelo escolhe entre alternativas prováveis e pode escolher caminhos diferentes a cada execução, principalmente com temperatura alta. Em contexto clínico, variar a resposta diante da mesma pergunta é um problema: a decisão precisa ser reproduzível.

12. O que significa uma resposta ser determinística?

Significa que, com a mesma pergunta e o mesmo acervo, o resultado é sempre o mesmo. Isso é essencial para auditoria: se dois profissionais fazem a mesma consulta em dias diferentes, eles precisam poder comparar o resultado sem suspeitar do sistema.

13. O que é treinamento de um modelo?

É o processo de mostrar bilhões de exemplos ao modelo e ajustar seus números internos até ele aprender os padrões da linguagem. É caro, demora semanas ou meses e só acontece de tempos em tempos.

14. O que é ajuste fino?

É um treinamento complementar, menor, feito para especializar um modelo genérico em um assunto. Ele ensina o estilo e o comportamento da tarefa, mas não garante que o modelo saiba cada caso do hospital.

15. A inferência muda o modelo?

Não. Na inferência o modelo é somente leitura. Ele não aprende com a conversa, não aprende com o prontuário que recebeu e não memoriza a consulta. Tudo o que o modelo sabe sobre aquele caso está no que foi enviado no momento.

16. O que é um sistema RAG?

RAG é a sigla inglesa para 'geração aumentada por recuperação'. Na prática, é um sistema que, antes de o modelo responder, procura nos documentos da instituição os trechos mais relevantes para a pergunta e os entrega ao modelo junto com a pergunta. O modelo responde olhando para esses trechos.

17. Por que RAG surgiu?

Porque treinar o modelo de novo a cada documento novo é caro e lento, e porque um modelo genérico não conhece a rotina interna de um hospital específico. Com RAG, a informação fica organizada fora do modelo e é injetada na hora.

18. Qual é a diferença entre o modelo e o RAG?

O modelo é o motor de escrita. O RAG é a biblioteca organizada que entrega a ele o material certo. O modelo sabe escrever; o RAG decide o que ele vai ler antes de escrever.

19. O que é embedding, ou vetor semântico?

É a transformação de um texto em uma lista de números que representa o sentido dele. Dois textos com sentidos parecidos produzem listas parecidas. Assim o computador consegue comparar sentido, e não apenas palavras.

20. O que é busca semântica?

É a busca que compara sentido. Se você pergunta 'pressão alta com IECA', ela encontra um prontuário que fala de 'hipertensão arterial em uso de enalapril', mesmo sem repetir as palavras da pergunta.

21. Qual é a diferença entre busca por palavra e busca semântica?

A busca por palavra procura a palavra exata: se o prontuário usa sinônimo, ela erra. A busca semântica compara sentido: ela encontra o caso mesmo com palavras diferentes, mas pode trazer coisas parecidas que não servem.

22. O que é busca híbrida?

É a combinação das duas buscas anteriores. Ela usa palavra exata para não perder precisão e sentido para não perder cobertura. É o padrão mais confiável em acervo clínico.

23. O que é recuperação de informação?

É a área que estuda como encontrar os documentos certos em uma coleção grande. É a ciência por trás de qualquer busca, inclusive a busca de prontuários.

24. O que é similaridade de cosseno?

É uma forma de medir o quanto dois vetores apontam na mesma direção. O valor vai de -1 a 1; quanto mais perto de 1, mais parecido é o sentido. É a nota de semelhança que aparece na tela do Tracer-MED.

25. Uma nota de semelhança alta significa que o prontuário é o caso certo?

Não por si só. Ela diz que o sentido do texto está próximo do sentido da pergunta. O caso certo precisa ser decidido pelo profissional; a nota serve para ordenar os candidatos e para registrar por que um documento entrou ou ficou de fora.

26. O que é um índice vetorial?

É uma estrutura que guarda os vetores de todos os documentos de forma organizada para permitir comparação rápida. Sem ele, comparar a pergunta com cada prontuário do hospital seria uma operação lenta.

27. O que é chunk, ou trecho?

É um pedaço de documento indexado separadamente. Um prontuário longo pode virar vários trechos. O cuidado é não cortar o sentido do meio, porque um trecho cortado perde precisão.

28. Por que a divisão em trechos importa?

Porque o motor encontra apenas o que está no trecho. Se a informação que importa ficou em outro pedaço, ela não aparece. Uma divisão ruim gera resultados que parecem incompletos.

29. O que é top-k?

É a quantidade de resultados que o sistema devolve. Top-6 significa os seis mais próximos da pergunta. O resto não entra na resposta, o que torna essencial saber por que o resto foi descartado.

30. O que é reranking?

É uma segunda passada de comparação, mais lenta e mais rígida, sobre os candidatos trazidos na primeira passada. Ela reordena poucos documentos para que os mais relevantes fiquem no topo.

31. Por que existem duas etapas de recuperação?

A primeira etapa precisa ser muito rápida para lidar com todo o acervo. A segunda pode ser mais cara porque olha apenas para poucos candidatos. Duas etapas dão velocidade e qualidade ao mesmo tempo.

32. O que é filtro exato?

É um filtro por um campo preciso, como CID-10, departamento ou data. Ele não depende de parecença: ou o registro tem aquele valor, ou não tem. É o que permite dizer 'só quero Cardiologia dos últimos 90 dias'.

33. O que é post-filter?

É aplicar o filtro exato depois de recuperar os candidatos. É simples e rápido, mas exige cuidado: se o filtro for aplicado só no fim, a quantidade devolvida pode ficar abaixo do esperado.

34. O que é um falso positivo na busca?

É quando o sistema traz um documento que parece relevante pela nota, mas não serve para a pergunta. O profissional percebe e ignora. O risco é ele passar batido em uma leitura apressada.

35. O que é um falso negativo na busca?

É quando o documento relevante existe no acervo, mas não aparece no resultado. É o erro mais perigoso em saúde, porque o profissional não vê o que não recebeu e pode concluir errado sem perceber.

36. Qual é o erro mais grave em uma busca clínica?

O falso negativo. Trazer um registro a mais incomoda; não trazer o registro que decide o caso pode levar a uma conduta inadequada. Por isso a prioridade é provar que nada relevante ficou de fora.

37. O que é recall, em linguagem simples?

É a fração de tudo o que era relevante que o sistema de fato encontrou. Recall alto significa que pouco ficou de fora. Recall baixo significa que o sistema perdeu material importante.

38. O que é precisão, em linguagem simples?

É a fração do que o sistema devolveu que realmente era relevante. Precisão alta significa pouca coisa inútil na lista. Precisão e recall disputam espaço: ganhar um costuma custar o outro.

39. O que é nota de corte?

É o valor mínimo de semelhança para entrar na resposta. Na prática, é a nota do último resultado selecionado. Tudo o que dá menos que isso fica de fora, e é justamente aí que a proveniência do descarte importa.

40. O que é um acervo documental?

É o conjunto de documentos que o sistema pode consultar. No Tracer-MED, o acervo de um hospital é formado pelos prontuários daquela instituição, com marcação de data, CID, departamento e versão.

41. O que é um prontuário eletrônico?

É o registro digital da assistência de um paciente: consultas, evoluções, exames, prescrições. É a matéria-prima de qualquer busca clínica de verdade.

42. Por que dados de saúde exigem cuidado especial?

Porque são dados pessoais sensíveis. A LGPD trata a saúde como categoria especial. Qualquer sistema que consulte prontuário precisa de base legal, finalidade definida, controle de acesso e rastro de quem consultou.

43. O que é anonimização?

É retirar ou mascarar os dados que identificam a pessoa. Não é o mesmo que pseudonimização.

44. O que é pseudonimização?

É trocar os identificadores diretos por códigos, mantendo a possibilidade de reidentificação controlada. É o modelo mais comum em índice de busca clínica.

45. O que é isolamento multi-tenant?

É garantir que os dados de uma instituição nunca apareçam para outra. Cada hospital é um tenant separado. O isolamento precisa estar no motor de busca, e não apenas na tela.

46. Por que o isolamento na tela não basta?

Porque esconder o botão não impede o acesso. Se o motor tem o acervo de todos juntos, um erro de configuração ou uma consulta direta pode vazar dados. O isolamento tem de acontecer na camada que guarda e procura os documentos.

47. O que é governança de IA?

É o conjunto de regras, registros e controles que permite usar IA de forma responsável e demonstrável: quem usou, para quê, com qual dado, com qual resultado e com qual conferência.

48. O que é rastreabilidade?

É a capacidade de reconstruir o caminho de uma decisão. Em busca clínica, significa saber qual pergunta foi feita, qual acervo foi consultado, qual versão do índice estava ativa e quais documentos entraram ou ficaram de fora.

49. O que é auditabilidade?

É a qualidade de poder ser conferido por alguém de fora. Um sistema auditável deixa provas independentes do próprio sistema: o resultado não precisa ser aceito por confiança.

50. O que é uma trilha de auditoria?

É um registro cronológico, encadeado e à prova de adulteração dos eventos do sistema: consulta feita, relatório emitido, evidência assinada. Cada evento aponta para o anterior, então alterar um quebra a corrente.

2 Dados, treino, limites e governança

De onde vem a informação, por que o modelo erra, o que significa prova, trilha de auditoria e conformidade. É o bloco que sustenta a conversa com compliance e jurídico.

PERGUNTAS 51 A 100.

51. O que é WORM?

WORM significa 'grava uma vez, lê muitas'. É um registro que não pode ser editado nem apagado depois de criado. É a base do histórico de evidências do Tracer-MED.

52. O que é assinatura digital?

É um selo criptográfico que prova que um documento foi gerado por quem diz ter gerado e que não foi alterado depois. As tecnologias citadas no sistema são Ed25519 para a assinatura e SHA3-256 para o encadeamento por hash.

53. O que é hash?

É como uma impressão digital de um conteúdo. Se uma vírgula mudar, o hash muda completamente. É isso que permite provar que um registro não foi alterado.

54. O que é fingerprint de uma busca?

É a impressão digital daquela combinação: pergunta, tenants, filtros, versão e parâmetros. Ela permite provar que aquela resposta foi gerada por aquela configuração específica.

55. O que é um LLM no contexto da saúde?

É o modelo de linguagem usado para redigir, resumir ou conferir texto clínico. Ele é útil e conveniente, mas não é fonte de verdade: ele é o redator, não o arquivo.

56. Por que os hospitais querem usar IA generativa?

Porque ela reduz tempo em tarefas de texto: resumir um caso, redigir uma evolução, organizar pedidos, responder a perguntas de profissionais. O ganho de tempo é real; a confiança precisa ser construída.

57. Por que a pressa em adotar IA é um risco?

Porque a resposta de um modelo é fluente mesmo quando está errada. Sem uma camada que traga o material certo e registre o que foi usado, o hospital não consegue mostrar depois como aquela decisão foi tomada.

58. O que é hallucination rate ou taxa de alucinação?

É a proporção de afirmações inventadas ou não sustentadas em uma amostra de respostas. Medir essa taxa é o primeiro passo para decidir se um modelo pode entrar em determinado fluxo de trabalho.

59. Uma taxa de alucinação de uma marca serve para a minha instituição?

Serve como referência geral, não como garantia local. O comportamento do modelo depende do tipo de pergunta, do acervo e do processo da instituição. O que vale é a conferência sobre o seu próprio material.

60. O que é ground truth, ou referência de verdade?

É a resposta considerada correta por especialistas, usada para comparar o que o sistema produziu. Sem referência, não há medida: só impressão.

61. O que é avaliação por amostragem?

É sortear uma parte das respostas e revisar com especialista. É o método prático para acompanhar qualidade em produção sem revisar tudo.

62. Por que é impossível provar que um texto gerado por modelo está 'cl clinicamente correto'?

Porque 'cl clinicamente correto' depende de julgamento, contexto e diretriz. O que pode ser provado é algo mais simples e verificável: que cada frase pode ser levada a um documento que estava presente no material enviado ao modelo.

63. O que é contexto fechado?

É quando o modelo recebe apenas um conjunto definido e registrado de documentos para responder. Fora desse conjunto, ele não tinha material. Isso transforma a conferência em uma pergunta objetiva: a frase está ou não está sustentada por algo que foi entregue?

64. Por que contexto fechado ajuda a auditabilidade?

Porque o universo que o modelo podia usar fica registrado. Se a frase menciona um prontuário que não estava no contexto, a conferência acusa. Se estava, é possível verificar a semelhança entre frase e documento.

65. O que é prompt?

É o texto que se envia ao modelo: a instrução, a pergunta e o material de apoio. A qualidade do material de apoio define boa parte da qualidade da resposta.

66. O que é engenharia de prompt?

É a prática de escrever instruções melhores para obter respostas melhores. Ela ajuda, mas não substitui o material: bom prompt com acervo errado continua gerando resposta fraca.

67. O que é citação obrigatória?

É a regra de que cada afirmação precisa vir acompanhada do documento que a sustenta, no formato usado pelo Tracer-MED: o número do prontuário entre colchetes, como [PR-000412]. Sem citação, a frase não pode ser conferida.

68. Por que citação obrigatória é mais forte que pedir para o modelo 'ser honesto'?

Porque honestidade é uma instrução que o modelo pode seguir ou não. Citação é um objeto verificável: ou existe o identificador, ou não existe. A conferência trabalha sobre o que é verificável.

69. O que é verificador de afirmações?

É a camada que pega a resposta, divide em frases e testa cada frase contra o material que estava disponível. O resultado é um veredito por frase, não um selo geral.

70. O que o verificador consegue e o que ele não consegue provar?

Ele prova que a frase pode ser levada a um documento presente no contexto. Ele não prova que o modelo interpretou bem o documento, que o documento é representativo do caso ou que a conduta é a melhor. Esses pontos precisam ficar declarados como não provados.

71. O que é confiança calibrada?

É quando a confiança que o modelo expressa corresponde à chance real de acertar. É um problema difícil e, no estado atual, não resolvido de forma confiável. Por isso a conferência trata a confiança como algo a declarar, não a usar como prova.

72. O que é viés em um modelo de linguagem?

É a tendência de reproduzir padrões exagerados ou distorcidos presentes nos dados de treino ou no material enviado. Em saúde, isso pode significar sub-representar grupos ou repetir condutas antigas de forma acrítica.

73. Por que 'a IA é neutra' é uma afirmação errada?

Porque o modelo aprende com textos que já têm padrões e desigualdades. Ele pode amplificá-los. A neutralidade não é uma propriedade automática: é um trabalho de avaliação.

74. O que é privacidade diferencial?

É uma técnica que adiciona ruído controlado às respostas estatísticas para que não seja possível identificar uma pessoa específica. É tema de pesquisa em acervos clínicos.

75. O que é retenção de dados?

É o tempo que um sistema guarda um dado. Precisar de prova não significa guardar tudo para sempre: o rastro de auditoria deve guardar o necessário para conferência, e não o conteúdo clínico sem necessidade.

76. O que é um log de consulta

É o registro de que uma consulta foi feita, por quem, quando, com qual pergunta e com qual resultado. Ele permite reconstruir o atendimento e responder a questionamentos de compliance.

77. O que é um relatório de busca auditável?

É o documento que transforma uma consulta em prova: pergunta, instituição, acervo, modelo, resultados, descartes justificados, recibo e assinatura. É o que o Tracer-MED gera em PDF e imprime em folha A4.

78. Qual é a diferença entre prova e promessa em IA?

Promessa é a afirmação de que o sistema funciona. Prova é um objeto que outra pessoa consegue conferir sem confiar no fornecedor. O Tracer-MED foi desenhado para transformar o máximo de promessas em provas, e para declarar com clareza o que continua sendo promessa.

79. O que é um limite declarado?

É a parte do funcionamento que o sistema admite não provar. Declarar limites é parte da credibilidade: um sistema que afirma provar tudo não deve ser levado a sério.

80. O que é 'suficiência de contexto'?

É ter, dentro do material enviado, tudo o que é necessário para responder com segurança. Se falta informação, o modelo tende a preencher o vazio. A garantia de completude ataca justamente o risco de o material chegar incompleto.

81. O que é memória de contexto em um assistente?

É a quantidade de informação que o assistente carrega entre mensagens. Ela é limitada. Em prontuários longos, o limite obriga a escolher o que enviar, e a escolha precisa ser defensável.

82. O que é uma consulta em linguagem natural?

É perguntar do jeito que se fala: 'quais pacientes hipertensos com IECA melhoraram a pressão?'. O sistema traduz a pergunta em comparação de sentido e devolve os registros mais próximos.

83. Por que buscar em português é diferente de buscar em inglês?

Porque as ferramentas de sentido mais usadas foram treinadas predominantemente em inglês. O português exige modelos e testes próprios, sobretudo em vocabulário clínico brasileiro.

84. O que é vocabulário clínico controlado?

É uma lista padronizada de termos, como o CID-10 para diagnóstico e o TUSS para procedimentos. Ele reduz ambiguidade: 'hipertensão' sempre significa a mesma coisa no sistema.

85. O que é CID-10?

É a classificação internacional de doenças usada para codificar diagnósticos. No Tracer-MED, o CID-10 é um filtro exato e uma marca de organização do acervo.

86. O que é SNOMED CT?

É uma terminologia clínica mais detalhada que o CID-10, usada em prontuários eletrônicos e interoperabilidade. Ela descreve achados, procedimentos e situações clínicas.

87. O que é HL7 FHIR?

É um padrão de troca de informações de saúde entre sistemas. Ele define como prontuário, paciente e exames são transportados de um sistema para outro, sem que cada hospital invente seu próprio formato.

88. O que é interoperabilidade?

É a capacidade de sistemas diferentes trocarem informação de forma compreensível. Sem ela, o acervo de um hospital fica preso dentro de um único software.

89. O que é um data lake clínico?

É um repositório que guarda grandes volumes de dados de saúde em formato bruto, para uso analítico. Ele não é a mesma coisa que um índice de busca: um guarda, o outro encontra.

90. O que é governança do dado?

É o conjunto de regras sobre quem pode criar, ver, mudar e descartar um dado. Ela define o que a IA pode consultar.

91. O que é qualidade de dado em prontuário?

É o grau de completude, padronização e confiabilidade do registro. Prontuário com campo vazio, abreviação caseira ou data errada compromete qualquer busca, por melhor que seja o motor.

92. Por que a qualidade do índice importa tanto quanto o modelo?

Porque o modelo só responde sobre o que recebe. Índice desatualizado ou mal organizado entrega material errado, e o modelo erra com aparência de acerto.

93. O que é envelhecimento de índice?

É a defasagem entre o acervo real e o índice. Se um exame novo não foi indexado, ele simplesmente não aparece na busca. Por isso o relatório registra a versão e a data do índice.

94. O que é reindexação?

É refazer o índice para incluir documentos novos ou corrigidos. É operação de rotina em qualquer sistema de busca sério.

95. O que é um benchmark de recuperação?

É um conjunto de perguntas com respostas conhecidas usado para medir sistemas de busca e comparar versões. Sem benchmark, não há como dizer se a nova versão melhorou ou piorou.

96. O que é Recall@10?

É a proporção de casos em que o documento relevante aparece entre os dez primeiros resultados. É uma medida comum para comparar motores de busca.

97. O que é um corpus de teste?

É a coleção de documentos usada para avaliar o sistema antes de colocá-lo em uso com dados reais.

98. O que é validação externa de um sistema?

É quando alguém fora do fornecedor testa o sistema em cenário próprio. Ela vale mais do que qualquer número publicado pelo fabricante.

99. O que é um erro de omissão em relatório clínico?

É deixar de registrar algo que deveria estar no documento. Em auditoria, omitir é tão grave quanto registrar errado.

100. O que é reprodutibilidade?

É a capacidade de outra pessoa chegar ao mesmo resultado usando os mesmos dados e a mesma configuração. Ela é a base de qualquer afirmação científica ou auditoria.

3 O mercado de inferência médica e a análise comercial

Quem compra, quem vende, o que se mede, quais objeções aparecem e como uma conversa comercial honesta se conduz. Inclui o custo de conferir, o piloto e o ROI.

PERGUNTAS 101 A 200.

101. O que é o mercado de inferência médica?

É o conjunto de produtos e serviços que usam modelos para apoiar decisões e tarefas clínicas: organizar prontuário, resumir caso, sugerir conduta, apoiar triagem, redigir laudo, responder a profissional. Ele envolve fornecedores, hospitais, operadoras, clínicas e áreas de compliance.

102. Quem compra inferência médica no mercado brasileiro?

Hospitais de médio e grande porte, grandes clínicas, operadoras de saúde, laboratórios, redes de diagnóstico e secretarias de saúde. Em geral, a compra é decidida entre a área assistencial e a área de tecnologia, com parecer de compliance e jurídico.

103. Quem são os fornecedores típicos?

Três grupos: os donos do modelo (OpenAI, Google, Anthropic), os prontuários eletrônicos que embutem IA no que já vendem e os especialistas em camada de dados, que trazem indexação, busca e governança sobre o acervo da instituição. O Tracer-MED atua no terceiro grupo.

104. Por que o mercado cresceu tanto nos últimos anos?

Porque a IA generativa reduziu o esforço de lidar com texto livre, que é o formato predominante em saúde. Resumir um caso, organizar uma evolução e responder perguntas passaram a ser viáveis em escala, o que abriu uma classe de produto que antes dependia de regras manuais.

105. Como se mede o tamanho desse mercado?

Principalmente por projeções de duas famílias: mercado de governança de IA e mercado de IA em saúde. São estimativas de consultorias privadas, variam entre fontes e devem ser lidas como ordem de grandeza, não como número exato.

106. Por que os números de mercado importam para quem compra um RAG clínico?

Importam como contexto, não como argumento de compra. O que decide a adoção é o custo do erro, o tempo de conferência e a capacidade de mostrar a auditoria. Uma projeção de mercado grande não prova que o sistema funciona no seu hospital.

107. Quais são os segmentos do mercado de IA em saúde?

Diagnóstico por imagem, apoio à decisão clínica, documentação e codificação, gestão de leitos e escalas, recepção e jornada do paciente, faturamento e auditoria de contas, e governança de dados. O Tracer-MED se posiciona em documentação e apoio à decisão com prova.

108. Onde o mercado brasileiro está mais maduro?

Em imagem e em documentação. São áreas com métrica clara e retorno visível. Já a governança de IA aplicada a texto clínico ainda é pouco atendida, e é exatamente aí que a busca auditável faz diferença.

109. O que é apoio à decisão clínica?

É qualquer recurso que ajuda o profissional a decidir: protocolo, cálculo de risco, alerta de interação, resumo do histórico, busca de casos parecidos. O Tracer-MED não decide nada: ele entrega o caso relevante com a prova de que não deixou outro de fora.

110. Qual é a diferença entre IA diagnóstica e IA de apoio documental?

A diagnóstica emite um parecer sobre o paciente (por exemplo, aponta uma lesão em exame). A documental ajuda a encontrar, resumir e organizar o que já existe. A documental tem risco menor de erro clínico direto e exige prova de não omissão.

111. O que é codificação assistida?

É usar IA para sugerir códigos de CID ou de procedimento a partir do texto do atendimento. Ela acelera o faturamento, mas erra fácil: precisa de conferência humana e de registro do que foi sugerido.

112. Qual é o impacto financeiro de uma codificação ruim?

Glosa, atraso de recebimento, retrabalho e risco de questionamento em auditoria. É uma das razões pelas quais sistemas de codificação tem sido os primeiros a comprar governança.

113. O que é glosa?

É a recusa parcial ou total de pagamento por um plano ou convênio, geralmente por inconsistência entre o que foi cobrado e o que foi documentado. Uma busca que prova o que estava no prontuário ajuda a contestar glosa.

114. Por que hospitais procuram IA para reduzir tempo?

Porque parte do tempo do profissional vai para localizar informação dispersa: exames em um sistema, evolução em outro, prescrição em um terceiro. Buscar em vários lugares e reconstruir o caso é um custo invisível e recorrente.

115. Qual é o principal medo de um gestor de hospital ao adotar IA?

Um erro clínico com consequência jurídica que ele não consiga explicar. Por isso o critério de compra sai do "o modelo é bom" e vai para "o que eu consigo mostrar depois".

116. O que compliance costuma pedir antes de aprovar IA?

Base legal para tratar dados, controle de acesso por perfil, isolamento entre instituições, registro de quem consultou o quê, retenção definida, plano de remoção de dados e clareza sobre quem responde pelo resultado.

117. O que o jurídico de um hospital costuma pedir?

Que o sistema não invente informação, que a fonte seja identificável, que a responsabilidade profissional continue clara e que o contrato delimite o papel do fornecedor como apoio, e não como substituto do médico.

118. O que a área médica costuma pedir?

Velocidade e confiança. O médico aceita rápido um resultado que ele entende e confere, e recusa qualquer coisa que o obrigue a sair da tela para verificar.

119. O que a TI costuma pedir?

Baixo acoplamento: que a solução não obrigue troca de prontuário, tenha padrões abertos (como HL7 FHIR), seja observável em produção e não crie uma segunda base de dados sem governança.

120. Por que os três compradores raramente concordam de imediato?

Porque medem coisas diferentes: o médico mede tempo e confiança, a TI mede integração e risco técnico, o compliance mede prova e exposição. Projetos de IA que passam apenas em um dos três tendem a travar na etapa de aprovação.

121. Qual é o teste mais comum que hospital aplica a um fornecedor de IA?

O piloto com uma pergunta simulada e um caso real de controle. Eles perguntam algo cuja resposta o time médico já sabe e verificam se o sistema traz o prontuário correta e se nada relevante foi omitido.

122. Como um fornecedor de busca clínica se diferencia de um chatbot?

Um chatbot responde; uma busca clínica auditável entrega o material com fonte e comprova a cobertura. A diferença aparece quando aparece errado: no chatbot, não há como provar; na busca auditável, há um relatório com o que entrou e o que ficou de fora.

123. Por que o Fornecedor de modelo puro não resolve o problema do hospital?

Porque o fornecedor do modelo não conhece o acervo do hospital, não controla o isolamento entre tenants, não sabe o que é relevante naquele serviço e não emite relatório de auditoria daquela instituição. Ele entrega o motor de escrita.

124. Por que um prontuário eletrônico com IA embutida nem sempre basta?

Porque ele conhece apenas os dados que passa pelo próprio sistema, com qualidade de cadastro que tem, e porque a IA tende a vir acoplada à decisão de compra do prontuário. Um público independente consegue provar o que incluiu e o que não incluiu.

125. O que é o custo de inferência e por que ele pesa?

É o custo de executar o modelo por consulta. Ele cresce com o tamanho do material enviado. Enviar o acervo inteiro em toda consulta é impraticável em custo e em tempo; por isso existe a etapa de recuperação.

126. Qual é a economia do RAG no gasto com modelo?

Mesmo uma consulta de alguns milhares de tokens em vez de um acervo inteiro, e uma redução do custo por consulta de uma ordem de grandeza. Em escala hospitalar, a diferença decide a viabilidade do projeto.

127. O que é custo total de propriedade de um sistema de IA clínico?

Não é só licença e inferência. Entra indexação, integração com prontuário, treinamento da equipe, auditoria periódica, manutenção e o custo do retrabalho quando o resultado não é confiável.

128. Como um hospital deveria comparar dois fornecedores de RAG?

Pelo mesmo caso, com o mesmo acervo, medindo três coisas: encontrou o prontuário certo?, deixou algo relevante de fora?, conseguiu provar o que fez? Qualquer comparação baseada só em demonstração comercial é fraca.

129. O que é uma prova de conceito, ou PoC, em saúde?

É um teste curto, com escopo fechado e critério de sucesso acordado antes. Um PoC bom define o que será medido, quem revisa e qual é o número que faz o projeto avançar ou parar.

130. O que é piloto clínico controlado?

É roda o sistema em um serviço por um período, comparando resultados com um grupo de controle ou com a prática anterior. Em saúde, é o caminho mais honesto para afirmar ganho.

131. O que se deve medir num piloto de busca clínica?

Tempo até encontrar o caso, taxa de prontuários com fonte correta, taxa de casos em que algo relevante foi omitido, aceitação do médico e volume de consultas auditáveis emitidas.

132. O que é desfecho de qualidade em saúde?

É o resultado que interessa ao paciente e à instituição: tempo de permanência, reinternação, adesão, erro evitado, tempo de espera. A IA só se sustenta em um hospital quando aparece ligada a algum desses números.

133. Como a busca auditável pode não reduzir reinternação?

Ela não interfere diretamente no tratamento. Ela melhora a informação disponível no momento da decisão. O ganho vem de combinação com protocolo e conduta. Prometer redução de reinternação sem estudo é exagero comercial.

134. O que é ROI de um sistema de busca clínica?

Normalmente vem de horas poupadas na reconstrução de caso, tempo de alta, redução de glosa e redução de retrabalho de auditoria. Nenhum desses números é automático: depende do desenho da implantação.

135. Por que é difícil calcular o ROI de governança de IA?

Porque o valor está em evitar um evento raro e caro (um erro não provado, uma glosa grande, uma falha de auditoria). O gestor precisa separar economia mensurável de risco evitado.

136. O que é risco evitado?

É o custo que não aconteceu. Elege-se contratar porque o cenário ruim é caro, mas contabilizar isso é difícil. A comparação correta é com o custo de um evento de auditoria malsucedido.

137. O que é responsabilidade profissional em uso de IA?

O ato médico continua sendo do médico. O sistema é apoio. Sistemas bem desenhados deixam isso explícito na tela, no relatório e no contrato, para que ninguém atribua ao modelo uma decisão que continua humana.

138. O que é o CFM e por que ele aparece nas discussões?

É o Conselho Federal de Medicina. A Resolução CFM 1.821/2007 trata da guarda e do manuseio do prontuário. Ela entra quando o sistema gera, guarda ou consulta documento clínico.

139. O que a LGPD exige de um sistema de busca em prontuário?

Finalidade definida, base legal (a tutela da saúde costuma enquadrar o uso assistencial), minimização do dado, segurança, controle de acesso, registro de tratamento e definição de retenção.

140. O que é base legal para tratar dado de saúde?

É a justificativa jurídica que autoriza uso. Para assistência, é comum a tutela da saúde por profissional do setor, conforme a LGPD. Uso para treinar modelo com finalidade diferente exige outro enquadramento.

141. Por que usar o prontuário para assistência não autoriza treinar modelo com ele?

Porque a finalidade muda. A tutela da saúde cobre o cuidado do paciente, não necessariamente a construção de produto. Separar as finalidades é obrigação, não formalidade.

142. O que é HIPAA e por que é citada em hospital brasileiro?

É a regra norte-americana de privacidade em saúde. Ela aparece no Brasil como referencia de padrão quando o hospital atende ou integra grupo internacional, ou quando o fornecedor quer demonstrar maturidade.

143. O que é GDPR e o que é o Artigo 9?

É o regulamento europeu de proteção de dados. O Artigo 9 trata de categorias especiais, inclusive dados de saúde, e exige salvaguardas adicionais. Ele aparece em grupos com operação na Europa.

144. Uma bandeira de conformidade na tela significa que o hospital está em conformidade?

Não. Ela indica apoiar um requisito. A conformidade é do processo do hospital, não do software. Por isso a bandeira deve vir acompanhada da explicação do que ela cobre.

145. O que é consentimento e quando ele é necessário?

É a manifestação do titular sobre o uso do dado. Em assistência, a base legal costuma ser a tutela da saúde e não o consentimento. Em pesquisa ou produto, o consentimento e a submissão a comitê de ética podem ser necessários.

146. O que é comitê de ética em pesquisa?

É a instância que avalia projetos que usam dados de pacientes para pesquisa. Qualquer uso secundário de prontuário em estudo precisa passar por ele.

147. O que é termo de uso assistencial?

É o documento que registra a finalidade do uso do acervo na assistência. Ele apoia a transparência e a auditoria, mas não substitui a base legal.

148. O que é auditoria de acesso?

É a revisão de quem consultou o que. Ela identifica acesso fora do horário do plantão, consulta de paciente de outro serviço e padrão incomum de busca.

149. O que é o princípio do menor privilégio?

É dar a cada perfil apenas o acesso necessário. É a razão pela qual o Tracer-MED tem perfis distintos: médico assistencial vê o texto clínico; auditoria vê metadados e relatório; TI vê o funcionamento técnico.

150. Por que o auditor não deveria ver o texto clínico?

Porque o trabalho dele é verificar o processo, não o conteúdo do paciente. Restringir o texto reduz exposição e mantém a auditoria focada na prova.

151. O que é mascaramento de dado?

É ocultar parte de um campo, como mostrar só os últimos dígitos de um documento ou as iniciais do nome. Permite conferência sem expor o dado inteiro na tela, no relatório ou no material enviado ao modelo.

152. O que é minimização de dado?

É usar apenas o dado necessário para a finalidade. Enviar ao modelo o prontuário inteiro quando três trechos bastam é desperdício e exposição.

153. Por que enviar contexto enxuto reduz risco legal?

Porque reduz exposição do dado e facilita demonstrar finalidade. Um contexto registrado e pequeno é mais fácil de defender do que um envio do acervo inteiro.

154. O que é contrato de tratamento de dados?

É o documento entre hospital e fornecedor que define o que é tratado, com qual finalidade, por quanto tempo e com qual segurança. Ele é exigível em auditoria.

155. O que é suboperador em cadeia de IA?

É o fornecedor do fornecedor, como a API do modelo. Ele recebe dado do hospital, então precisa estar declarado no contrato, com finalidade, retenção e local de processamento, do mesmo jeito que o fornecedor principal.

156. Por que um hospital não pode usar API pública com prontuário real sem análise?

Porque o dado sai do controle da instituição. Antes, é preciso verificar contrato, retenção, local de processamento e se o fornecedor treina com os dados enviados.

157. O que é retenção zero do modelo?

É a garantia contratual e técnica de que o provedor não guarda o conteúdo enviado para treino ou uso posterior. É um requisito comum em contratos de saúde.

158. O que é modelo local e por que hospital considera?

É rodar o modelo dentro da infraestrutura do hospital. Ganha controle de dado e isolamento; perde qualidade em relação aos maiores modelos e exige investimento em hardware e equipe.

159. Qual é a vantagem de um sistema independente do modelo escolhido?

O hospital não fica preso. Ele pode começar com um modelo, migrar para outro e manter o acervo, o índice, a busca e a prova. A prova precisa sobreviver à troca de modelo.

160. O que significa ser agnóstico de modelo?

Significa funcionar com diferentes provedores via interface padrão, como a compatível com a API OpenAI. O Tracer-MED fala com OpenAI, Gemini, Qwen, modelos locais via Ollama e gateways.

161. Por que trocar de modelo é mais fácil que trocar de índice?

Porque o modelo é substituível por uma chamada; o índice guarda o trabalho de organização do acervo e o histórico de versões. A parte que fica é a camada de dados.

162. O que é um gateway de modelos?

É uma camada que padroniza a chamada a vários provedores, como o LiteLLM. Ela simplifica troca, controle de custo e registro de uso em um único ponto.

163. O que é observabilidade de IA?

É conseguir enxergar o que o sistema está fazendo em produção: consultas, latência, erros, custo por consulta, taxa de conferência sem apoio e alertas de mudança de comportamento.

164. O que é deriva, ou drift, em um sistema de IA?

É quando o comportamento muda com o tempo, por atualização do modelo, mudança de acervo ou mudança de prática clínica. Sem monitorar, a qualidade cai sem aviso.

165. O que é avaliação contínua em produção?

É amostrar consultas reais, revisar com especialista e acompanhar os números ao longo do tempo. É o único modo de saber se a garantia continua valendo depois da instalação.

166. O que é o custo de conferir?

É o tempo que médico ou auditor gasta para validar uma resposta. Sistemas ruins escondem esse custo. Bons sistemas o reduzem, mostrando a fonte ao lado da frase.

167. Por que o custo de conferir é o gargalo comercial?

Porque sem ele não há adoção. O médico não vai conferir linha por linha em toda consulta. Se conferir for caro, ele deixa de usar ou usa no automático. A prova precisa ser barata de ler.

168. O que é uma resposta auditável?

É uma resposta em que cada afirmação pode ser levada a uma fonte que estava disponível e em que o conjunto de fontes está documentado com data e versão. Ela não precisa ser verdadeira para ser auditável, mas precisa ser verificável.

169. Qual é a diferença entre auditável e sempre correta?

Auditável é o que se pode conferir. Sempre correta é o que não se pode provar facilmente. O Tracer-MED escolhe o caminho da auditabilidade porque ele pode ser demonstrado.

170. O que é o argumento comercial central do Tracer-MED?

O hospital quer usar IA, mas precisa defensável. O Tracer-MED entrega o material certo ao modelo, prova que nada relevante ficou de fora e deixa o relatório assinado para auditoria. Ele é a camada de confiança entre o prontuário e o modelo.

171. Quem é o concorrente real do Tracer-MED?

Na prática, a planilha e o processo manual de conferência. A empresa não compete com o GPT; compete com a forma como o hospital hoje reconstrói um caso e com a falta de registro de quem consultou o quê.

172. Como apresentar o Tracer-MED a um diretor clínico?

Pelo caso, não pela tecnologia: "o médico pergunta do jeito que fala, recebe os prontuários com fonte e o hospital fica com o documento que prova o que foi consultado". Tecnologia só entra se ele perguntar.

173. Como apresentar ao compliance?

Pelo conjunto: isolamento por instituição, perfis de acesso, trilha encadeada e assinada, relatório com data, versão do acervo e nome do emissor. É o que ele pode levar a uma auditoria.

174. Como apresentar à TI?

Pelos padrões: front-end estático, API compatível com OpenAI, servidor proxy que mantém a chave fora do navegador, geração de PDF própria e sem dependência externa pesada. Ele avalia risco de integração, e a integração é barata.

175. Qual objeção comercial aparece primeiro?

"Nós já temos busca no prontuário." A resposta é mostrar a diferença entre encontrar por palavra e encontrar com prova de não omissão. Uma demo com dois campos parece igual; a diferença aparece quando você pergunta o que ficou de fora.

176. Qual é a segunda objeção?

"E se o modelo errar?" A resposta é separar entrada e saída: a entrada é provada pelo motor; a saída é conferida frase a frase contra o contexto. O que não pode ser provado fica declarado como não provado.

177. Qual é a terceira?

"Como isso se integra ao nosso prontuário?" A resposta é por padrão aberto. O índice lê o acervo via exportação ou HL7 FHIR e o resultado volta para a tela do profissional.

178. Qual é a quarta?

"Quanto custa?" O custo tem três partes: licença, integração e custo de inferência por consulta. A comparação honesta é com o tempo hoje gasto para reconstruir um caso e com o custo de uma glosa contestada com sucesso.

179. Qual é a quinta?

"E o piloto?" O piloto precisa de critério antes, acervo real de uma unidade e revisão por amostra. Sem isso, o piloto vira demonstração de entusiasmo.

180. O que o cliente nunca deveria ouvir de um fornecedor de IA clínica?

"O modelo nunca erra" ou "pode confiar". Nenhum modelo é assim. Um fornecedor sério trata o erro como parte do desenho e mostra como ele é detectado.

181. O que é uma demonstração honesta?

É uma demo que mostra também o limite: a resposta sem apoio, o caso em que nada foi encontrado, o comportamento sem chave de API. Demos só com caminho feliz não sobrevivem ao piloto.

182. Por que muitos projetos de IA em saúde morrem depois do piloto?

Porque o piloto mede entusiasmo e a produção mede consistência. Sem monitoramento, sem integração real e sem dono do processo, o uso cai quando a equipe volta à rotina.

183. O que é um dono do processo?

É a pessoa na instituição responsável por manter o sistema vivo: atualizar acervo, revisar amostras, responder a alertas. Sistemas de IA sem dono viram projeto piloto abandonado.

184. O que é treinamento de equipe em sistemas de IA?

É ensinar o uso correto: como perguntar, como ler o relatório, como reconhecer uma frase sem apoio, o que fazer quando o resultado parece estranho. Adoção depende disso mais do que da interface.

185. Como medir aceitação em adoção?

Por uso recorrente, não por número de acessos. Se o médico usa toda semana, se o relatório é anexado a processos de auditoria e se o volume cresce sem cobrança, há aceitação real.

186. O que é expansão natural de um sistema como esse?

Começa em um serviço (cardiologia, triagem), com consultas simples. Depois cresce para outros serviços, para auditoria de contas, para governança de dados e para outros produtos da linha (GOV, JUS, GAP) com o mesmo motor.

187. Por que a linha Winnex tem vários produtos sobre o mesmo motor?

Porque a pergunta é sempre a mesma: dado um acervo regulado, encontre o que importa e prove que nada ficou de fora. Muda o acervo (prontuário, contrato, processo, apólice) e o nome do produto.

188. O que é Tracer-GOV?

É a aplicação da mesma lógica a acervo público: contratos, licitações e transparência. A diferença é regulatória e de vocabulário, não de raciocínio de prova.

189. O que é Tracer-JUS?

É a aplicação a processos e jurisprudência: encontrar precedente relevante com prova de que nada importante foi omitido. É a área em que falso negativo custa em processo.

190. O que é Tracer-GAP?

É a aplicação a documentos financeiros e de seguros. A pergunta é localizar cláusula e evidência em apólice e contrato, com rastreabilidade.

191. O que essa arquitetura permite oferecer a um grupo de saúde?

Um mesmo padrão de prova em vários acervos: assistência, faturamento, contratos e jurídico. Isso é o que sustenta a venda para grandes grupos, não apenas para um serviço.

192. O que é um caso de uso de entrada recomendado para o piloto?

Busca de casos parecidos por sintoma ou CID em um único serviço, com conferência em amostra. É simples, tem valor visível para o médico e não toca em decisão direta de tratamento.

193. Qual caso deve ficar fora do primeiro piloto?

Qualquer uso que substitua conduta, dose ou diagnóstico definitivo. O sistema apoia e informa, mas não decide, e o piloto deve começar onde isso é evidente para todos.

194. O que é argumento de venda baseado em prova, não em feature?

É demonstrar o funcionamento em vez de listar funcionalidades. Em vez de "temos IA", mostrar: "fizemos a pergunta, ele trouxe seis prontuários com fonte, descartou os outros com justificativa calculada e emitiu o relatório assinado".

195. O que o cliente compra de fato?

Compra a resposta para uma dor de confiança. Ele não compra vetores nem o motor; compra a possibilidade de usar IA sem precisar pedir que alguém acredite nele.

196. O que diferencia prova de marketing de prova de auditoria?

A prova de auditoria pode ser refeita por terceiro com os mesmos dados e a mesma configuração. O marketing não se refaz; é apenas afirmação.

197. O que é a pergunta de ouro que um comprador informado faz?

Mostre-me este mecanismo funcionando no meu próprio acervo, com um caso que eu escolho, e me diga em qual passo ele deixa de provar. Um fornecedor que responde isso ganha confiança que nenhum deck constrói.

198. O que é a pergunta de ouro seguinte?

E quando o modelo de linguagem erra? A resposta correta separa entrada provada e saída conferida, e declara o que não se prova. Quem evita essa pergunta está escondendo o ponto fraco.

199. Por que a prova importa mais agora do que há cinco anos?

Porque a IA generativa tornou fácil gerar texto convincente e difícil verificar origem. Quando todo mundo consegue produzir resposta fluente, o diferencial migra para quem consegue provar de onde veio.

200. O que é reputação de dados em IA clínica?

É a confiança acumulada de que o sistema mostra fonte, registra uso e não infla resultado. Ela se perde com uma única promessa exagerada demonstrada em piloto.

4 RAG na prática, a stack e a mecânica do Tracer-MED

Como a busca acontece, como o acervo é obtido, como o relevante é encontrado, como o descarte é provado, como a resposta do modelo é conferida e o que o relatório garante.

PERGUNTAS 201 A 300.

201. O que é RAG, em uma frase?

RAG é procurar primeiro e responder depois. Antes de o modelo escrever, o sistema busca nos documentos da instituição os trechos mais próximos da pergunta e entrega esses trechos ao modelo junto com a pergunta.

202. Por que o nome fala em "aumentar" a geração?

Porque o material recuperado aumenta o que o modelo tem para trabalhar. Sem ele, o modelo responde só com o que aprendeu no treino; com ele, responde olhando para o acervo atual da instituição.

203. Quais são as etapas de um RAG?

Quatro: preparar o acervo, transformar pergunta em vetor, recuperar os candidatos mais próximos e entregar esses candidatos ao modelo com a pergunta. O Tracer-MED acrescenta a prova de cada etapa.

204. O que é a etapa de preparo do acervo?

É ler os documentos, dividir em trechos, transformar cada trecho em vetor e guardar tudo em um índice. É a etapa mais longa e a que mais influencia a qualidade de todo o resto.

205. O que é a etapa de consulta?

É transformar a pergunta em vetor, comparar com todos os vetores do índice e separar os mais próximos. É a etapa que decide o que o modelo vai ver.

206. O que é a etapa de geração?

É o modelo escrevendo a resposta a partir da pergunta e dos trechos entregues. Ela pode produzir bom ou mau texto mesmo com material bom; por isso existe a conferência frase a frase.

207. Dê três exemplos de RAG em uso comum.

Atendimento ao cliente com base de artigos, suporte interno com documentação da empresa e assistente jurídico com contratos. Em todos, o modelo não sabe o conteúdo: ele lê na hora.

208. Por que esses exemplos funcionam?

Porque o acervo é escrito com cuidado, é relativamente pequeno e a resposta errada custa tempo, não saúde. Em saúde, o acervo é ruidoso e o erro tem consequência maior, o que exige mais prova.

209. O que é RAG de busca simples por palavra?

É o mais barato: sem vetores, apenas índice de texto. Funciona para documentos com vocabulário padronizado e falha quando o profissional usa sinônimo ou linguagem livre.

210. O que é RAG de busca semântica?

É o que compara sentido por vetores. Encontra o caso mesmo com palavras diferentes, mas pode trazer parecidos que não servem e deixa de registrar a justificativa dos descartes.

211. O que é RAG híbrido?

É a combinação das duas buscas com uma etapa de reordenação. É o desenho mais comum em acervo clínico porque concilia cobertura e precisão.

212. O que é RAG com reordenação?

É recuperar muitos candidatos rápido e depois reordenar os melhores com um comparador mais rígido. Aumenta qualidade no topo da lista ao custo de tempo.

213. O que é RAG com filtros exatos?

É usar metadados como CID, departamento e data como filtro antes ou depois da busca. Garante que a resposta respeite restrições objetivas, não parecidas.

214. O que é RAG com prova?

É o desenho do Tracer-MED: além de recuperar, calcular para cada documento descartado um limite máximo de semelhança e mostrar que esse limite fica abaixo da nota de corte. Assim o descarte deixa de ser presumido e passa a ser demonstrado.

215. Qual é a diferença entre RAG comum e RAG com prova?

O RAG comum responde "estes são os mais parecidos". O RAG com prova responde "estes são os mais parecidos e, destes outros, posso demonstrar que nenhum poderia estar na resposta".

216. Por que o RAG com prova é mais importante em saúde que em comércio?

Porque em comércio o pior caso é uma resposta incompleta; em saúde, o pior caso é a ausência do registro que muda a conduta. Provar o descarte é proteger o paciente.

217. O que é um agente RAG?

É um RAG que decide sozinho os passos: reformular a pergunta, escolher onde buscar, consultar várias fontes, validar e responder. Ele amplia cobertura ao custo de mais complexidade e mais dificuldade de auditoria.

218. Por que o Tracer-MED não deixa o agente responder livremente?

Porque liberdade sem prova reproduz o problema. No Tracer-MED, o agente pode planejar, mas a resposta só avança sobre material provado e cada etapa é registrada.

219. O que são os agentes listados na demo?

São funções de trabalho: triagem, evidência, FAQ, encaminhamento e auditoria. Cada um tem um papel, um limite e uma dependência. Não são um enxame de decisões independentes.

220. O que o Agente de Triagem faz?

Transforma a pergunta clínica em resultados com garantia de cobertura. Ele não responde sozinho: entrega os prontuários e a prova para o médico decidir.

221. O que o Agente de Evidência faz?

Assina e encadeia cada consulta no histórico e emite o relatório de busca auditável. Ele não busca; ele registra e prova a busca que já aconteceu.

222. O que o Agente de FAQ faz?

Responde perguntas frequentes a partir de documentos aprovados pela instituição. Ele trabalha com material curado, não com prontuário livre, e por isso não entra no acervo clínico do paciente.

223. O que o Agente de Encaminhamento faz?

Lê o motivo do atendimento e sugere departamento e protocolo de encaminhamento, com base nos protocolos atualizados pela regulação. Ele roteia o caso; não decide conduta e não substitui a avaliação de quem recebe o paciente.

224. O que o Agente de Auditoria faz?

Varre a trilha e confirma que cada registro bate com o anterior; se algo foi alterado, acusa. É o vigia do próprio histórico: a garantia de imutabilidade só vale se for verificada de tempos em tempos.

225. Por que os agentes têm a mesma prova do médico?

Porque sem isso o agente vira caixa-preta. A mesma regra que vale para a resposta do médico vale para o plano de um agente: nada avança sobre material descartado sem justificativa.

226. O que é o motor Madhava?

É o motor de recuperação da linha Winnex. Ele varre o acervo, ordena por semelhança e, para cada documento que não entra, calcula o máximo de semelhança que ele poderia atingir. É esse número que sustenta a frase "a exclusão está demonstrada".

227. O que significa dizer que o Madhava é um motor com garantia?

Significa que ele devolve, além da resposta, a demonstração de que os documentos fora da resposta não poderiam estar nela. A garantia é sobre cobertura, não sobre qualidade da escrita.

228. O que é o Normalize na stack?

É a camada que cuida da qualidade numérica dos vetores antes da busca. Se os vetores estão mal formados, a comparação de semelhança fica distorcida. O Normalize garante que o acervo tem qualidade para ser medido.

229. Qual é o trabalho de cada peça, em uma frase?

O Normalize prepara a medida; o Madhava mede e prova que nada relevante ficou de fora; o Tracer assina, encadeia e emite o relatório; o produto entrega isso na tela do profissional.

230. O que é o Tracer, dentro da stack?

É a camada de evidência, com os comandos de registro e verificação do histórico: `make_worm_record` para gravar e `verify_record` para conferir. É o que transforma busca em prova guardada.

231. O que é o Tracer-MED, dentro dessa stack?

É o produto da área de saúde: interface de busca para o médico, filtros clínicos, relatório por prontuário, perfis de acesso e isolamento por instituição, tudo apoiado no Madhava e registrado pelo Tracer.

232. O que é a demo visual?

É a réplica navegável do produto, com dados fictícios, que permite ver funcionando a busca, a garantia, o relatório e a trilha sem instalar nada no hospital. É material de demonstração, não o sistema em produção.

233. O que a demo NÃO prova?

Ela não prova o motor em acervo real. Os números da demo são ilustrativos, e o próprio documento declara isso. Quem quiser prova de escala precisa de instalação com o acervo da instituição.

234. Por que uma demo é necessária mesmo quando existe a tecnologia?

Porque a conversa sobre IA clínica é abstrata até alguém ver os próprios casos funcionando. A demo transforma a discussão em experiência e antecipa as objeções.

235. O que significa dizer que a exclusão foi comprovada?

Significa que, para aquele documento, o sistema calculou o melhor que ele poderia ter feito e esse melhor ficou abaixo da nota do último resultado da resposta. Mesmo no cenário mais favorável, ele não entraria.

236. O que é o "resíduo ortogonal" e por que ele aparece?

É a parte do documento que não é explicada pela direção principal da consulta. Quanto menor o resíduo, mais confiável é a estimativa de semelhança máxima. Ele é o termo técnico que garante o limite.

237. Por que uma desigualdade matemática serve como prova?

Porque ela não depende de opinião: se o máximo possível de um documento é menor do que a nota de corte, não existe configuração em que ele entraria. É uma afirmação que se pode conferir refazendo o cálculo.

238. O que é a nota de corte na prática?

É a semelhança do último resultado que entrou. Ela é a fronteira: quem está acima entra, quem está comprovadamente abaixo fica de fora com justificativa registrada.

239. O que é a tabela de descartes?

É a lista dos prontuários que não entraram, cada um com a semelhança máxima, a nota de corte e o motivo da exclusão. Ela é o coração da transparência do Tracer-MED.

240. Por que mostrar o que ficou de fora é mais valioso do que mostrar só o que entrou?

Porque o erro perigoso é a omissão. Quem vê apenas a resposta não sabe se faltou algo. Quem vê a resposta e a lista de descartes justificados consegue julgar a completude da busca.

241. O que é o banner de garantia na tela?

É o resumo em linguagem direta: quantos documentos foram conferidos, quantos foram descartados com justificativa, quantas falhas na conferência e quanto tempo levou. Ele traduz a prova para quem não vai abrir a tabela.

242. O que significa "falhas na conferência: 0"?

Significa que nenhum documento descartado apareceu depois com semelhança acima da nota de corte. É a conferência da própria garantia. Em instalação real, esse número é medido na execução.

243. Como o Tracer-MED encontra o que é relevante?

Comparando o sentido da pergunta com o sentido de cada trecho do acervo, refinando com uma segunda passada e aplicando filtros exatos de CID, departamento e período. O relevante é o que fica acima da nota de corte depois desses passos.

244. Como a pergunta do médico vira busca?

O texto é transformado em vetor pelo mesmo modelo que indexou o acervo. A comparação acontece entre vetores, não entre palavras. Por isso a pergunta em linguagem livre funciona.

245. O que acontece se a pergunta for vaga?

O sistema traz candidatos por proximidade de sentido, que podem ser numerosos e pouco precisos. A resposta continua auditável, mas menos útil. Perguntas específicas melhoram o resultado.

246. O que acontece se a pergunta não tiver resposta no acervo?

O sistema não deve inventar. Ele mostra que nenhum registro passou da nota de corte e registra isso. Saber que não há registro é informação clínica valiosa.

247. Como o Tracer-MED obtém o conjunto de dados?

Do próprio hospital: dos prontuários eletrônicos da instituição, exportados por integração ou por padrão de interoperabilidade. O acervo não vem de fora; ele é o registro assistencial da casa.

248. O que é a versão do acervo?

É a identificação de qual conjunto de prontuários estava no índice naquele momento. Sem essa marca, um relatório de auditoria não se sustenta: não se sabe sobre qual acervo a busca foi feita.

249. Por que o período do acervo é registrado?

Porque define o alcance temporal da busca. "Últimos 90 dias" e "todo o histórico" são buscas diferentes. O relatório precisa dizer qual foi feita.

250. O que é pseudonimização no índice?

É manter nome, CPF e CNS fora do vetor, guardando apenas o que identifica o caso no sistema clínico. O motor compara sentido, e sentido não exige identificação direta.

251. Por que o modelo de embedding aparece no relatório?

Porque a nota de semelhança depende dele. Trocar o modelo muda as notas e a ordem. Registrar qual modelo leu o acervo é registrar as condições da medida, como um laudo registra o equipamento usado.

252. O que é a dimensão do vetor, como o 1024 que aparece na tela?

É o tamanho da lista de números que representa cada trecho. Dimensão maior costuma capturar mais nuances, ao custo de mais memória e mais tempo de comparação.

253. O que é a etapa 1 e a etapa 2 da busca?

A etapa 1 traz candidatos rápido, olhando o acervo inteiro. A etapa 2 reordena um grupo menor com um comparador mais rígido. Os números 64 e 128 na tela são os tamanhos dessas etapas.

254. Por que a busca precisa ser rápida e provada ao mesmo tempo?

Porque o médico está no atendimento. Uma prova que chega em dez minutos não muda a conduta. O desenho do sistema assume que o custo da prova precisa caber no tempo da consulta.

255. Como o Tracer-MED avalia as respostas de outros modelos?

Aceita o texto de qualquer provedor compatível, identifica o que é citação de prontuário e compara cada frase com o material que estava no contexto. O resultado é frase a frase: com apoio, sem apoio ou não verificável.

256. O que é contexto fechado na conferência?

É a regra de que o modelo recebeu apenas os prontuários recuperados. Assim, a pergunta objetiva é: a frase pode ser levada a um desses prontuários? Se não pode, ela não é sustentada pelo acervo.

257. Por que a conferência trabalha frase por frase?

Porque um selo geral esconde a frase errada no meio do texto correto. Frase por frase, o profissional vê exatamente onde a resposta perdeu o apoio.

258. O que é "frase sem apoio"?

É a frase que não encontra no contexto um prontuário que a sustente. Ela aparece marcada, e a recomendação é revisão humana antes de usar.

259. O que é "não verificável"?

É a frase que a conferência não consegue testar, por falta de elemento comparável. Declarar isso é mais honesto do que transformar incerteza em selo verde.

260. Como a conferência mede apoio entre frase e prontuário?

Comparando o conteúdo informativo da frase com o conteúdo do prontuário citado: título, resumo e marcas clínicas. O resultado é uma proporção que permite classificar a frase como apoiada ou não.

261. Por que a citação de prontuário ajuda o verificador?

Porque restringe a busca do apoio. Se o modelo cita o prontuário certo, a comparação é direta; se cita outro, a discrepância aparece. Sem citação, a conferência fica mais fraca.

262. O que a conferência NÃO prova?

Ela não prova relevância clínica do caso, calibração de confiança do modelo e fidelidade interpretativa. Esses três pontos são declarados como não provados — e essa declaração é parte do produto.

263. Por que a conferência é feita na saída se o motor já prova a entrada?

Porque a entrada provada não impede a saída inventada. O motor garante que o material certo chegou; a conferência garante que o texto não foi além do material.

264. Como o Tracer-MED ajuda o GPT, o Gemini ou o Claude a responder melhor?

Dando a eles o material certo e só ele. O modelo deixa de depender do que memorizou no treino e passa a responder sobre o caso real daquela instituição, com o contexto reduzido e registrado.

265. Por que o modelo responde melhor com contexto do que sem contexto?

Porque a resposta deixa de ser uma média de tudo o que ele viu e passa a ser uma leitura do caso concreto. O modelo deixa de adivinhar padrão geral e passa a organizar evidência específica.

266. O que é "mesma pergunta, mesma conduta" em um hospital?

É a consistência de resposta entre profissionais e turnos. Com acervo e regra registrados, dois médicos que fazem a mesma pergunta chegam ao mesmo conjunto de prontuários e podem discutir a conduta sobre a mesma base.

267. O que é variabilidade de resposta entre modelos?

É a diferença de texto entre GPT, Gemini e Claude diante da mesma pergunta e do mesmo contexto. Ela é esperada e não compromete a auditoria, desde que o material enviado seja o mesmo e registrado.

268. O que o Tracer-MED controla e o que ele não controla?

Controla o acervo, a busca, a nota de corte, o registro e o relatório. Não controla a redação do modelo, se o modelo escolheu bem a conduta e se a conduta é clinicamente a melhor. Essa separação é o centro do produto.

269. Por que a resposta de um modelo pode ser diferente em um hospital e em outro?

Porque o acervo é diferente, a rotina é diferente, a codificação é diferente e a população atendida é diferente. Mesmo caso, regiões diferentes, práticas diferentes. Injetar o prontuário correto aproxima a resposta do que aquela instituição de fato faz e registra.

270. Por que a mesma doença pode exigir conduta diferente em lugares diferentes?

Porque disponibilidade de exame, perfil de resistência, protocolo local, rede de referência e prevalência variam. Ler o registro local evita recomendar algo que faz sentido em teoria e não existe na prática daquele serviço.

271. O que é orientação sem registro local?

É a recomendação genérica. Ela pode ser correta em tese e inútil ali. Por isso a busca local importa: ela traz o que a equipe já fez naquele contexto.

272. Como o Tracer-MED usa o histórico do próprio hospital?

Trazendo casos parecidos já atendidos naquela instituição, com data, CID, departamento e conduta registrada. O médico compara o caso atual com a experiência documentada da casa.

273. O que é aprendizado com a prática local?

É usar o acervo do próprio serviço como referência. Não é treinar o modelo: é consultar o que o serviço já registrou. O modelo continua o mesmo; o material muda.

274. Por que isso importa para a segurança do paciente?

Porque uma decisão apoiada no registro da própria equipe é mais fácil de conferir por quem vai executar. A base está visível e pode ser questionada no mesmo idioma do serviço.

275. O que é personalização por instituição?

É ter acervo, protocolo e relatório próprios de cada hospital. O sistema é o mesmo; o conteúdo é de quem contrata. O isolamento garante que a personalização não vaze.

276. Qual é o caminho completo, do pedido do médico ao relatório, em oito passos?

Um: o médico digita a pergunta. Dois: a pergunta vira vetor. Três: o motor procura no índice do tenant. Quatro: aplica filtros e reordena. Cinco: calcula o descarte dos demais. Seis: entrega os prontuários e a remissão da prova. Sete: o modelo, se configurado, redige com contexto fechado. Oito: o sistema confere a saída e registra tudo na trilha, permitindo emitir o relatório.

277. O que o relatório de busca auditável contém?

Identificação da instituição e do emissor, a pergunta, os filtros, o acervo com versão e período, o modelo, a garantia de completude, os prontuários apresentados, os descartes com justificativa, a validação da resposta e a trilha de evidência, com recibo e assinatura.

278. Por que o relatório tem recibo e QR de verificação?

Porque o relatório circula fora do sistema. O QR carrega o endereço de verificação com tenant e recibo, e permite conferir a origem do documento. É a ponte entre o papel e o registro.

279. O que é um relatório por prontuário?

É o documento individual de um registro: de onde veio, com qual modelo foi lido, por que entrou na resposta, qual a semelhança e qual o selo de verificação. Ele permite anexar a defesa de um caso específico ao prontuário.

280. Por que o relatório é por prontuário e por consulta ao mesmo tempo?

Porque o hospital precisa dos dois níveis: o documento da consulta para auditoria e o documento do caso para o processo daquele paciente. Os dois saem do mesmo conjunto de dados.

281. O que é a trilha de auditoria WORM?

É a sequência de eventos encadeados em que cada registro contém a impressão digital do anterior. Alterar um evento quebra a corrente e fica evidente. Ela é a memória que não pode ser reescrita.

282. Por que encadear por hash impede a adulteração?

Porque cada evento aponta para o conteúdo do anterior. Mudar um registro antigo exige recalcular todos os seguintes, e a assinatura do último deixa de conferir.

283. O que é o fingerprint de uma consulta?

É a impressão digital da combinação de pergunta, instituição, filtros, período e quantidade de resultados. Ela identifica a consulta de forma única e aparece no relatório.

284. O que é a assinatura Ed25519 mencionada no sistema?

É o método criptográfico com que a evidência é assinada. Ela prova autoria e integridade. O recibo e o fingerprint dão a identidade da consulta; a assinatura dá a garantia de que ninguém alterou depois.

285. O que é SHA3-256?

É a função de hash usada no encadeamento. Ela produz uma impressão digital fixa para qualquer conteúdo; mudança de um caractere gera outra impressão.

286. Qual a diferença entre hash e assinatura?

O hash detecta mudança. A assinatura prova autoria. O primeiro responde "mudou?"; o segundo responde "quem gerou?". Os dois são necessários.

287. O que é um recibo de consulta?

É o identificador curto que liga o documento emitido à consulta registrada. É o que o paciente não vê, mas o auditor usa para localizar a evidência.

288. Como o Tracer-MED garante a prova?

Em três camadas: na busca, calculando o limite de cada descarte; na assinatura, registrando a consulta em cadeia imutável; no relatório, emitindo o documento com recibo, QR e assinatura do emissor. A prova não depende de examinar o código.

289. A prova depende de o modelo ser bom?

Não. A prova é sobre cobertura do acervo e integridade do registro. O modelo pode ser fraco e o relatório continuar honesto, porque a conferência da saída mostra o que ficou sem apoio.

290. A prova depende de o médico confiar no fornecedor?

Não deveria. O desenho permite ao hospital refazer a consulta e comparar com o relatório. Prova que exige confiança no fornecedor não é prova de auditoria.

291. O que é a seção "Detalhes técnicos da garantia"?

É o bloco recolhível com a fórmula, os parâmetros e a tabela de descartes. Ele fica fora do caminho do médico e disponível para auditoria e TI, que precisam reconstruir o raciocínio.

292. Por que o jargão fica escondido?

Porque o médico precisa de resposta, não de disciplina técnica. O jargão é necessário para a auditoria e por isso não é removido: é separado.

293. O que é a seção "Como esta resposta foi produzida"?

São quatro etapas em linguagem de usuário: de onde vieram os dados, qual modelo leu o acervo, como o prontuário foi encontrado e o que a resposta assegura. É explicação para quem não quer estudar o sistema.

294. O que é a seção "Conferir resposta de um modelo"?

É o verificador da saída: o texto do modelo é comparado com os prontuários que estavam no contexto. Ele existe porque a prova da entrada não impede a invenção na saída.

295. Por que a interface tem perfis de acesso?

Porque médico, auditoria e TI precisam de informações diferentes. O médico vê o texto clínico; o auditor vê metadados, garantia e relatório; a TI vê o funcionamento técnico. Cada um vê o necessário.

296. Como a demo trata o isolamento entre instituições?

Troca a sessão inteira: site, usuário, acervo, consultas e índice mudam juntos. O médico de um tenant não consegue ver consulta do outro porque o motor só o enxerga o próprio acervo. É a demonstração do isolamento na camada certa.

297. Como isso funciona com dados reais?

Cada instituição tem o próprio índice e o próprio conjunto de casos. O usuário autenticado é associado ao tenant antes de qualquer consulta. O motor não recebe parâmetro livre de tenant: recebe a sessão.

298. O que é prova de não omissão? E por que o Tracer-MED pode oferecê-la?

É a demonstração de que nenhum documento acima da nota de corte ficou fora da resposta. Ele pode oferecer porque calcula um limite, e não uma estimativa: para cada descartado, um máximo, e uma comparação. Sem cálculo de limite, não há prova de cobertura.

299. O que é o limite, em linguagem simples?

É a nota máxima que o documento alcançaria se tudo fosse favorável. Se até esse máximo fica abaixo da nota de corte, ele não entraria. É uma conta de pior caso.

300. Como o assistente de PDF e o relatório se apoiam?

O relatório é HTML em folha A4 isolada; o PDF é gerado no próprio navegador, sem bibliotecas externas. Isso torna o documento portátil: pode ser anexado a prontuário e a processo de auditoria, sem depender do fornecedor.