



Universidade do Minho
Escola de Engenharia

Ambientes Inteligentes para a Saúde
2024-2025

Trabalho individual

Inteligência Ambiente e Assistentes Virtuais

Assistente Virtual para Suporte Pré-Natal

Beatriz Amorim, PG56112

Mestrado em Engenharia Biomédica | Informática Médica

Conteúdo

1	Introdução	3
2	Domínio e contexto de aplicação	4
3	Framework Ollama	4
4	Resultados e análise crítica	6
4.1	Modelo Phi3	7
4.2	Modelo Llama	8
4.3	Modelo Gemma3	8
4.4	Modelo Mistral	9
5	Cenários Clínicos para avaliação do modelo	9
6	Reflexão crítica	12
6.1	Limitações identificadas e perspectiva de melhoria	12
7	Conclusão	14

1 Introdução

Este trabalho foi desenvolvido no âmbito da unidade curricular de Ambientes Inteligentes para a Saúde, com o objetivo de criar e implementar um assistente virtual de aconselhamento médico, recorrendo à *framework Ollama* e à utilização de modelos de linguagem de grande escala (LLM).

O assistente foi desenvolvido para responder a questões relacionadas com o acompanhamento durante a gravidez, disponibilizando informação útil e acessível para apoiar grávidas e futuros pais ao longo desta fase. Com esta abordagem, pretende-se explorar o potencial dos LLM na prestação de informação fiável e personalizada, que seja capaz de esclarecer dúvidas frequentes sobre o desenvolvimento do bebé, cuidados alimentares, medicação, sinais de alerta, estilo de vida, entre outros tópicos que sejam relevantes para este tema. Para a concretização deste objetivo, foi utilizado o *Ollama*, como ferramenta de execução dos modelos LLM.

Foram testados quatro modelos distintos - Phi, LLaMA, Gemma e o Mistral, com o intuito de comparar o tipo de respostas fornecidas por cada um.

Adicionalmente, foram ajustados diversos parâmetros no ficheiro *Modelfile* de cada modelo, com o objetivo de otimizar o seu comportamento e adaptar as respostas dadas ao contexto do aconselhamento pré-natal. Este *fine tuning* permitiu observar o impacto destas variáveis na qualidade do aconselhamento prestado por cada modelo.

2 Domínio e contexto de aplicação

Como já referido, o desenvolvimento do assistente virtual teve como base o contexto do acompanhamento durante a gravidez, focando-se na disponibilização de informação pertinente e acessível a futuros pais. Esta fase da vida é marcada por inúmeras dúvidas, preocupações e mudanças físicas e emocionais, o que leva muitas pessoas a procurarem frequentemente apoio e esclarecimentos sobre diversos aspetos relacionados com a gravidez.

A escolha deste contexto deve-se à sua relevância social e à elevada procura de informação sobre temas relacionados com o desenvolvimento do bebé e questões sobre a própria gravidez. Apesar da abundância de conteúdos disponíveis online, nem sempre estes estão apresentados de forma clara ou personalizada, o que pode originar confusão e ansiedade. [1]

O desenvolvimento do assistente virtual irá, assim, fornecer respostas rápidas, acessíveis e fundamentadas a perguntas frequentes por parte dos utilizadores durante a gravidez. Esta abordagem poderá contribuir, não só para o bem-estar das grávidas, como também pode servir como um complemento informativo ao acompanhamento clínico tradicional.

Entre as principais necessidades identificadas destacam-se:

- Clareza e acessibilidade na comunicação da informação;
- Personalização das respostas, tendo em conta o histórico e tempo de gravidez do utilizador;
- Apoio complementar ao acompanhamento médico, sem substituí-lo.

3 Framework Ollama

A *framework Ollama* foi utilizada como plataforma de execução e teste de modelos de LLM. Esta ferramenta facilita os testes com diferentes modelos, arquiteturas e parâmetros sem a necessidade de recorrer a outras aplicações. Através da utilização do ficheiro *Modelfile*, é possível personalizar o comportamento de cada modelo, ajustando alguns parâmetros fundamentais. [2]

Neste trabalho, os parâmetros utilizados e alterados foram os seguintes:

- **temperature**: altera o grau de criatividade ou aleatoriedade na geração de texto;
- **top_k**: define o número máximo de *tokens* mais prováveis a considerar em cada passo da resposta;
- **seed**: permite a reprodutibilidade dos resultados;

De modo a facilitar a interação entre o utilizador e os modelos, foi também utilizada a ferramenta *Open WebUI*, uma interface gráfica de código aberto desenvolvida especificamente para funcionar sobre a *framework Ollama*.

A integração da *Open WebUI* revelou-se particularmente útil na simulação de interações reais com utilizadores, sendo possível testar o desempenho dos modelos em contexto de utilização prática. Foram utilizados quatro modelos distintos:

- Phi3 (3.8B)
- LLaMA3.2 (3B)
- Gemma-3 (4B)
- Mistral (7B)

Importa salientar que o número de parâmetros de um modelo tem impacto direto na sua capacidade de compreensão e geração de texto. Modelos que tenham um menor número de parâmetros, como o LLaMA3.2 (3B) e o Phi3 (3.8B), tendem a ser mais rápidos e menos exigentes em termos de recursos computacionais. No entanto, esta leveza pode traduzir-se numa menor capacidade de guardar o contexto e de ter uma menor profundidade nas respostas, o que pode aumentar a probabilidade de ocorrerem erros científicos ou simplificações excessivas.

Em contrapartida, modelos maiores como o Mistral (7B) conseguem, em geral, elaborar respostas mais completas, coerentes e dentro do contexto. Ainda assim, este aumento de desempenho vem acompanhado de um custo computacional superior e maior tempo de espera pela resposta. Esta dualidade entre eficiência e qualidade de resposta foi um dos pontos avaliados no decorrer deste trabalho. Um exemplo claro desta limitação foi observado com os modelos Qwen e LLaMA 3.1 (8B), inicialmente considerados para os testes apresentados. Apesar do seu elevado potencial, ambos se revelaram demasiado exigentes em termos de recursos, apresentando tempos de resposta excessivamente longos para o equipamento disponível, o que levou à sua exclusão da análise final.

Quanto ao *prompt* do *Modelfile*, foi inicialmente utilizado o seguinte texto:

SYSTEM "You are a compassionate and knowledgeable prenatal assistant who supports pregnant individuals. You provide clear, kind, and evidence-based information about pregnancy stages, baby development, nutrition, medications, medical appointments, and exam preparation. You communicate using simple and accessible language.

If a question is unclear or missing important details (such as the pregnancy week, type of exam, or specific symptoms), ask for more information before responding. Do not guess or make medical diagnoses.

Always respond with empathy and reassurance. Your goal is to help the user feel informed, safe, and supported throughout their pregnancy.

If a question seems urgent or serious, gently recommend that the user contact a healthcare professional or go to the hospital instead of offering medical advice. "

Além deste *prompt*, foram dados vários exemplos de pergunta-resposta, de modo a que o modelo possa aprender o estilo de comunicação desejado, reconheça o contexto clínico e responda com informações precisas, empáticas e claras, ajustando-se às especificidades de cada caso apresentado. Estes exemplos servem para guiar o modelo na produção de respostas adequadas.

Durante os testes do modelo, o *prompt* foi ajustado para corrigir problemas específicos das respostas, resultando em versões diferentes do *prompt* aplicadas a modelos distintos. Entre as modificações, destacou-se a necessidade de evitar que o modelo abordasse a gravidez apenas como uma experiência extremamente positiva ou fizesse julgamentos de valor, procurando limitar o uso de expressões avaliativas como "wonderful", "beautiful", entre outros, ao mesmo tempo que mantinha um tom empático na conversa.

Be empathetic and supportive, but avoid subjective adjectives such as *wonderful*, *beautiful*, *exciting*, or *amazing* — do not express personal enthusiasm or value judgments about pregnancy.

Nem todos os modelos com esta restrição conseguiram seguir a indicação, o que é compreensível, dado que a empatia depende, em certa medida, do uso de alguns desses adjetivos. Contudo, essa medida revelou-se necessária, pois permitiu corrigir aqueles modelos que apresentavam uma visão excessivamente idealizada da gravidez.

4 Resultados e análise crítica

De modo a conseguir identificar diferenças no estilo, precisão clínica, capacidade de personalização, bem como avaliar o impacto dos parâmetros aplicados no *Modelfile*, foram escolhidas oito perguntas-chave sobre acompanhamento pré-natal:

1. What happens in week 12 of pregnancy?
2. Can I eat sushi while pregnant?
3. What vitamins should I take during the first trimester?
4. What are the best foods to reduce morning sickness?

5. What is the glucose tolerance test?
6. What should I ask at my 28-week check-up?
7. Is it safe to exercise in the third trimester?
8. How can I relieve back pain during pregnancy?

Para garantir uma comparação homogênea, aplicou-se a mesma sequência de perguntas a cada modelo analisado.

Cada conjunto de respostas foi analisado em termos de:

- **Clareza e estrutura:** Organização da informação e facilidade da compreensão do conteúdo.
- **Tom empático:** Uso de linguagem reconfortante e sensibilidade ao contexto emocional do utilizador.
- **Precisão clínica:** Alinhamento da informação com marcos de desenvolvimento fetal e recomendações de entidades médicas.
- **Detalhes e concisão:** Extensão das respostas e adequação ao contexto prático.

Com esta análise crítica, procura-se evidenciar não só as potencialidades de cada LLM no domínio do aconselhamento pré-natal, mas também as limitações que requerem supervisão clínica ou *fine-tuning*.

4.1 Modelo Phi3

De todos os modelos avaliados, o Phi foi o que apresentou uma menor precisão técnica, sobretudo no que diz respeito a marcos de desenvolvimento do bebé.

Em várias respostas, foram identificadas afirmações incorretas ou imprecisas, como a indicação de que a viabilidade fetal ocorre às 12 semanas ou que as contrações de Braxton-Hicks podem surgir no primeiro trimestre. Estes erros comprometem a fiabilidade das respostas, especialmente neste contexto onde a informação médica correta é essencial.

Na configuração sem parâmetros, o modelo apresentou um estilo direto e claro, com respostas organizadas, embora demasiado genéricas. O tom da conversa foi demasiado neutro ou impessoal, não incluindo elementos de empatia ou tentativas de personalizar a resposta consoante o contexto do utilizador.

Com a adição dos parâmetros, *temperature=0.4*, *top_k=30* e *seed=35*, o modelo tornou-se mais conversacional e empático, no entanto, as respostas tornaram-se excessivamente longas, por vezes redundantes, e houve um aumento de imprecisões técnicas.

4.2 Modelo Llama

O modelo LLaMA 3.2 (3B), demonstrou um desempenho superior ao Phi em termos de clareza e concisão, mas ainda apresentou limitações significativas que requerem atenção para a sua aplicação num contexto tão sensível quanto a gravidez.

Em geral, as dificuldades apresentadas, são semelhantes ao modelo anterior, porém com menor gravidade em alguns aspetos, o que o torna uma opção relativamente mais confiável, embora ainda longe de ser ideal para o aconselhamento pré-natal.

A diferença entre o modelo com parâmetros (*temperature 0.3, seed 35, top_k 40*) e sem parâmetros foi notável, com respostas mais estruturadas, concisas e objetivas na configuração com parâmetros, graças à redução da aleatoriedade e à limitação de escolhas lexicais, embora a precisão clínica permaneça limitada. Um aspeto menos positivo do parâmetro *temperature* definido como 0.3, foi o tom mais direto e impessoal. Porém, quando testado com uma *temperature* de 0.5, o modelo passou a apresentar exageros e imprecisões factuais. Um exemplo disso foi a resposta sobre o consumo de *sushi* durante a gravidez, onde indicou incorretamente o tubarão como opção alimentar segura, e classificou o bacalhau como peixe com níveis de mercúrio relevantes. Estes erros evidenciam que *temperatures* mais elevadas aumentam a aleatoriedade do modelo, comprometendo a exatidão das respostas.

4.3 Modelo Gemma3

O modelo Gemma destacou-se em relação aos dois últimos modelos. Embora o modelo tenha sido avaliado com diferentes parâmetros, aquele que melhor respondeu aos critérios definidos foi o executado apenas com a *seed*, sem parâmetros adicionais.

Esta comparação foi realizada utilizando os seguintes conjuntos de parâmetros: *top_k* = 45 e *temperature* = 0.5; *top_k* = 40 e *temperature* = 0.4.

Todas as configurações do modelo Gemma apresentaram pequenas imprecisões científicas, especialmente na descrição do exame de tolerância à glicose, com menção a horários incorretos para as colheitas de sangue, ou na questão do consumo de *sushi*, com explicações que podem ser consideradas vagas ou insuficientemente claras. Porém estas imprecisões são mínimas comparadas com os modelos anteriormente testados.

Com os parâmetros *temperature* = 0.5, *top_k* = 45 e *seed* = 35, o modelo apresentou respostas mais conversacionais e extensas, como seria de esperar. No entanto, esta configuração levou à inclusão de detalhes menos relevantes e à omissão de informações que seriam mais interessantes referir.

Por outro lado, com a *temperature* = 0.4 e o *top_k* = 40, o modelo, gerou respostas mais concisas, técnicas e precisas, mas menos práticas e abrangentes, carecendo de conselhos que seriam relevantes para a grávida.

O modelo que aparenta ter uma *performance* mais equilibrada foi o modelo sem a *temperature* e o *top_k*, uma vez que combinou clareza, detalhe e simplicidade. Foi também o que demonstrou mais empatia com linguagem afetuosa (ex: "your little one") e conselhos importantes, demonstrando atenção às necessidades emocionais e práticas.

4.4 Modelo Mistral

O modelo Mistral revelou-se excelente na vertente comunicativa e educativa, permitindo ao utilizador obter explicações claras, com uma linguagem natural e tom acolhedor.

A forma como organiza as respostas, recorrendo a listas e resumos breves, facilita a leitura e a retenção da informação. Porém, este foco na fluidez acaba por comprometer, em alguns casos, a profundidade técnica e a precisão clínica dos conteúdos. A adição de *temperature=0.3* e *top_k=40* trouxe mais objetividade, mas evidenciou ligeira perda de empatia e tornou as respostas mais secas e curtas.

5 Cenários Clínicos para avaliação do modelo

Para testar o desempenho e a relevância do assistente virtual desenvolvido, foram definidos vários perfis de pacientes grávidas com condições clínicas distintas. Cada perfil inclui dados demográficos e clínicos e serve como base para a colocação de uma pergunta relevante ao contexto. Estes cenários visam avaliar a capacidade do modelo em fornecer respostas empáticas, informadas e ajustadas ao contexto médico. O modelo utilizado para responder a estes cenários foi o *Gemma*.

Cenário 1:

- **Nome:** Beatriz Oliveira
- **Idade:** 22 anos
- **Mês de gestação:** 7.º mês (28 semanas)
- **Histórico médico:** Anemia leve diagnosticada no 4.º mês
- **Sintomas atuais:** Inchaço nas pernas, dores de cabeça frequentes
- **Observações:** Faz suplementação de ferro e vitaminas

Pergunta colocada ao modelo: *"Is it normal to feel swelling in my legs at this stage?"*

"Based on the information provided, you are 28 weeks pregnant and experiencing swelling in your legs. It's important to note that you have a history of mild anemia, diagnosed in your 4th month of pregnancy, and you are currently taking iron supplements and vitamins. Swelling, particularly in the legs and feet, is a common symptom during pregnancy due to hormonal changes and increased blood volume. However, given your existing anemia diagnosis, it's important to discuss this swelling with your healthcare provider to rule out any related issues and ensure proper management. Please consult your doctor for further evaluation and guidance."

Esta resposta demonstra uma compreensão razoável do contexto clínico e uma postura responsável. Dado que o *prompt* especifica claramente que o modelo não deve realizar diagnósticos como se fosse um profissional de saúde, e considerando que se trata de grávidas — um grupo particularmente sensível e vulnerável, tanto do ponto de vista físico como emocional — compreende-se a natureza mais genérica da resposta apresentada. Nestes casos, é fundamental adotar uma abordagem prudente, evitando transmitir falsas seguranças ou recomendações médicas potencialmente inadequadas.

Cenário 2:

- **Nome:** Inês Ferreira;
- **Idade:** 27 anos;
- **Mês de gestação:** 5.º mês (20 semanas);
- **Histórico médico:** Episódios de depressão recorrente desde os 18 anos. Atualmente em acompanhamento psicológico e psiquiátrico;
- **Sintomas atuais:** Dificuldades em dormir, preocupações constantes com a saúde do bebé;
- **Observações:** Está medicada com antidepressivo de categoria segura na gravidez. Faz sessões de terapia cognitivo-comportamental quinzenalmente.

Pergunta colocada ao modelo: *"How can I deal with the constant fear of something going wrong?"*

"It's understandable that you're experiencing a constant fear of something going wrong, especially given your history of recurrent depression. It's important to acknowledge and address these feelings. Given your current situation – being 5 months pregnant and experiencing symptoms of depression – seeking support is key.

Several strategies may be helpful:

- **Continue with your therapy:** Maintaining your cognitive-behavioral therapy (CBT) sessions is crucial for managing your anxieties.
- **Talk to your healthcare provider:** Discuss your fears openly with your doctor or psychiatrist. They can assess your concerns and potentially adjust your medication if needed.
- **Relaxation techniques:** Practices like deep breathing, meditation, or mindfulness can help reduce anxiety levels.
- **Support groups:** Connecting with other expectant mothers can provide a sense of community and shared experiences.

Remember, seeking professional help is a sign of strength, and it's a vital step in ensuring your well-being and the health of your baby."

A resposta do modelo é empática e adequada. Demonstra compreensão dos aspetos emocionais da gravidez e oferece sugestões práticas alinhadas com boas práticas clínicas, sem ultrapassar os limites éticos da atuação de um assistente virtual.

Cenário 3:

- **Nome:** Catarina Lopes
- **Idade:** 31 anos
- **Mês de gestação:** 6.º mês (24 semanas)
- **Histórico médico:** Cardiopatia congénita compensada, diagnosticada na adolescência. Em seguimento regular com cardiologia.
- **Sintomas atuais:** Cansaço frequente, ligeiras palpitações ao esforço
- **Observações:** Toma medicação beta-bloqueadora de forma controlada. Faz ecocardiograma trimestral para monitorização.

Pergunta colocada ao modelo: *"Will I need a C-section because of my heart condition?"*

"Based on the information you provided, Catarina, being 6 months pregnant with compensated congenital heart disease that is regularly monitored with echocardiograms, it's not possible to say for sure whether you will need a C-section. Your current symptoms, such as fatigue and mild palpitations, as well as your heart condition, are important factors that your medical team is closely following.

The need for a C-section depends on many factors during pregnancy, including the baby's position, your overall health, and any complications that might arise. This will be continuously assessed by your healthcare team, especially during your echocardiogram check-ups.

So, please continue to follow all medical advice and keep up with your regular monitoring to ensure the best care for you and your baby."

A resposta apresentada é clara e demonstra cautela e respeito pelos limites da informação disponível. O modelo evita assumir decisões clínicas e remete, corretamente, para o acompanhamento da equipa médica, o que é essencial em casos de patologias cardíacas durante a gravidez.

6 Reflexão crítica

A análise comparativa entre os diferentes modelos avaliados evidenciou a importância do *fine tuning* dos parâmetros e da seleção criteriosa do modelo para garantir respostas fiáveis, precisas e empáticas, que são cruciais num contexto sensível como a gravidez.

Entre os modelos avaliados, o modelo Gemma destacou-se como a solução mais equilibrada para um assistente virtual de aconselhamento pré-natal, conseguindo conjugar da melhor forma dois requisitos fundamentais: o rigor técnico indispensável nesta área sensível e a fluidez comunicativa necessária para o diálogo com grávidas.

Para além disso, o modelo mantém um tom empático e afetuoso — fundamental para estabelecer uma relação de confiança com a grávida — ao mesmo tempo em que organiza as respostas em estruturas mais didáticas, com listas e subtítulos que facilitam a leitura. Esta combinação de empatia, objetividade e consistência torna-o particularmente indicado para um assistente virtual de suporte pré-natal, capaz de fornecer orientações práticas, detalhadas e emocionalmente ajustadas às necessidades do utilizador.

6.1 Limitações identificadas e perspetiva de melhoria

Com isto, também foi possível avaliar o potencial e as limitações atuais da utilização de modelos de linguagem de grande escala (LLM) na área do aconselhamento pré-natal.

Em primeiro lugar, os modelos de linguagem utilizados foram treinados com dados genéricos, o que pode comprometer a precisão e a relevância das respostas no contexto específico do aconselhamento médico na gravidez. Além disso, apesar de serem capazes de gerar texto em linguagem corrente, os modelos não possuem uma verdadeira capacidade de interpretação clínica, o que limita a sua fiabilidade para casos complexos que exigem avaliação médica especializada.

Para superar estas limitações, especialmente no que respeita à adaptação e especialização dos modelos, seria recomendável implementar uma abordagem de *Retrieval-Augmented Generation* (RAG), que permitiria um *fine tuning* mais eficiente e contextualizado, integrando fontes de informação atualizadas e específicas, melhorando assim a precisão e a relevância das respostas fornecidas.

Verificou-se igualmente uma limitação ao nível dos recursos computacionais disponíveis, o que impossibilitou a execução de modelos mais robustos e a realização de testes em mais condições e cenários.

Por fim, destaca-se a necessidade de enquadrar estes assistentes virtuais como ferramentas complementares ao acompanhamento clínico tradicional, e não como substitutos. A responsabilidade ética no desenvolvimento e utilização destes sistemas deve ser uma prioridade, com ênfase na transparência das suas limitações e na recomendação explícita de contacto com profissionais de saúde em situações urgentes ou complexas.

7 Conclusão

Este trabalho explorou o potencial dos modelos de linguagem de grande escala (LLM) como ferramentas de apoio no aconselhamento pré-natal, demonstrando tanto as suas capacidades promissoras quanto as limitações significativas que ainda persistem. A análise comparativa dos modelos Phi, LLaMA, Gemma e Mistral revelou diferenças marcantes no seu desempenho, com o Gemma destacando-se como a opção mais equilibrada para esta aplicação específica.

O modelo Gemma mostrou-se particularmente eficaz na combinação de rigor técnico com comunicação acessível, conseguindo fornecer informações médicas precisas de forma clara e empática. Esta capacidade de adaptar o discurso às necessidades do utilizador, mantendo a precisão clínica, posiciona-o como uma ferramenta valiosa para educação na saúde materna. No entanto, os testes também evidenciaram existem limitações importantes nos modelos, quando confrontados com situações clínicas complexas ou requerem decisões médicas especializadas.

Os resultados reforçam a importância de enquadrar estes sistemas como complementos - nunca substitutos - do acompanhamento médico profissional.

No futuro este sistema deveria focar-se na integração com bases de conhecimento médico especializado, na implementação de arquiteturas mais sofisticadas como RAG (*Retrieval-Augmented Generation*), e na validação clínica rigorosa das suas respostas.

Assim, este trabalho confirma que os assistentes virtuais baseados em LLM, quando devidamente implementados e supervisionados, podem contribuir significativamente para melhorar o acesso a informação fidedigna, reduzir a ansiedade durante a gravidez e otimizar o tempo clínico para os casos que realmente exigem atenção especializada.

Referências

- [1] Femke Geusens and Alkistis Skalkidou. A two-wave survey study examining the impact of different sources of pregnancy information on pregnancy-related anxiety among Swedish women. *European Journal of Midwifery*, 9(January):1–11, 1 2025.
- [2] Ollama. [ollama/docs/modelfile.md](https://ollama.com/docs/modelfile) at main · ollama/ollama.