Instituto Superior
de Engenharia
Politécnico de Coimbra

DEPARTAMENTO DE ENGENHARIA QUÍMICA E BIOLÓGICA

Aplicabilidade da Inteligência Artificial numa Unidade de Cuidados Intensivos Neonatais da ULS Coimbra

Dissertação
Mestrado em Sistemas Avançados de Gestão da Saúde

Rui Jorge Simões Castelo

Orientador:

Prof. Doutor Mateus Daniel Almeida Mendes

INSTITUTO POLITÉCNICO
DE COIMBRA

INSTITUTO SUPERIOR
DE ENGENHARIA
DE COIMBRA

Coimbra, Dezembro de 2025

Aplicabilidade da Inteligência Artificial numa Unidade de Cuidados Intensivos Neonatais da ULS Coimbra

Trabalho desenvolvido no RCM^{2+}

Centro de Investigação em Ativos Físicos e Sistemas de Engenharia.

Resumo

A Inteligência Artificial (IA) afirma-se como uma ferramenta transformadora nos cuidados intensivos neonatais, com potencial para melhorar significativamente os resultados clínicos dos recém-nascidos mais vulneráveis. A sua capacidade de analisar grandes volumes de dados complexos, identificar padrões e fornecer informação em tempo real está a revolucionar a forma como se monitorizam, diagnosticam e tratam Recém-Nascidos (RN) prematuros.

Uma das áreas de aplicabilidade da IA em Unidades de Cuidados Intensivos Neonatais (UCIN) tem sido na monitorização de sinais vitais. Os algoritmos de IA analisam de forma contínua dados de sensores, como a frequência cardíaca, a frequência respiratória e a saturação de oxigénio, para detetar alterações subtis que podem indicar uma deterioração clínica. Assim, possibilita intervenções precoces, que podem prevenir eventos adversos graves. Essencialmente, têm sido publicados trabalhos sobre sépsis neonatal, displasia broncopulmonar e ventilação mecânica. A IA está também a ser utilizada para analisar imagens médicas, como ecografias cerebrais e cardíacas, para detetar anomalias e auxiliar no diagnóstico precoce de condições como hemorragia intraventricular e cardiopatias congénitas.

Neste Projeto, estudou-se a aplicabilidade da IA sob a forma de um Large Language Model (LLM) como parceiro na prestação de cuidados na UCIN. Mais especificamente, testou-se se um LLM pode ajudar na adequação da nutrição dos RN internados, ao analisar e propor ajustes nutricionais autonomamente. Para executar esta atividade, o LLM analisa os diários clínicos, extrai os dados clínicos relevantes, compara com standards e, por fim, caracteriza a adequação e propõe os ajustes necessários. Esta proposta do agente inteligente deve ser, por fim validada pelo clínico e eventualmente, aceite.

No decorrer do trabalho foi possível aferir que nem todos os LLMs são equivalentes, sendo que um deles (Gemma 3) revelou-se o mais adequado. O desempenho dos LLM foi bastante diverso, com modelos com mais parâmetros a obterem resultados inaceitáveis e outros LLM menos complexos com resultados ainda assim razoáveis. O tempo necessário para execução também foi variável, apesar de se ter privilegiado a segurança nos dados extraídos, com base em métricas universais.

Confirmou-se a capacidade de um LLM offline extrair dados clínicos e, após algum processamento adicional, gerar uma proposta de caracterização e ajuste nutricional.

Palavras-Chave:

Inteligência Artificial; Redes Neuronais; Neonatologia; Unidade de Cuidados Intensivos Neonatais; Nutrição; Large Language Models.

Abstract

Artificial Intelligence (AI) is establishing itself as a transformative tool in neonatal intensive care, with the potential to significantly improve clinical outcomes for the most vulnerable newborns. Its ability to analyze large volumes of complex data, identify patterns and provide real-time information is revolutionizing the way premature newborns are monitored, diagnosed and treated.

One of the areas of application for AI in NICUs has been in monitoring vital signs. AI algorithms continuously analyze data from sensors, such as heart rate, respiratory rate and oxygen saturation, to detect subtle changes that may indicate clinical deterioration. This enables early intervention, which can prevent serious adverse events. Most of the works published concern the impact on early diagnosis of neonatal sepsis, bronchopulmonary dysplasia, and optimizing mechanical ventilation. AI is also being used to analyze medical images, such as brain and heart ultrasounds, to detect abnormalities and aid in the early diagnosis of conditions such as intraventricular haemorrhage and congenital heart disease.

This project aims to study the applicability of AI in the form of an LLM as a partner in providing care in the NICU. More specifically, it aims to test whether an LLM can help to ensure adequate nutrition for hospitalized newborns by autonomously analyzing and proposing nutritional adjustments. To perform this activity, the LLM would have to analyze clinical records, extract relevant clinical data, compare it with standards and, finally, characterize its adequacy and propose the necessary adjustments. This proposal would ultimately be validated by the clinician and eventually accepted.

During the course of the study, it was possible to ascertain that not all LLMs are equivalent, with one of them (Gemma 3) proving to be the most suitable. The performance of the LLMs was quite diverse, with models with more parameters obtaining unacceptable results and other less complex LLMs still achieving reasonable results. The time required for execution was also variable, although priority was given to the security of the extracted data, based on universal metrics.

The ability of an offline LLM to extract clinical data and, after some additional processing, generate a proposal for nutritional characterization and adjustment has been thoroughly confirmed.

Keywords:

Artificial Intelligence; Neural Networks; Neonatology; Neonatal Intensive Care Unit; Nutrition; Large Language Models.

Epígrafe

"A Medicina é a ciência da incerteza e a arte da probabilidade."

Sir William Osler (1849-1919)

Dedicatória

Dedico este trabalho à minha família e a todas as famílias que depositam os seus recém-nascidos aos cuidados da equipa da qual faço parte.

Que estejamos sempre acima das suas expectativas!

Agradecimentos

Ao Professor Doutor e Orientador Mateus Daniel Almeida Mendes por toda a dedicação, pelo apoio incansável, acompanhamento e incentivo prestados ao longo da realização deste trabalho.

Uma palavra especial de apreço para o Professor Doutor Torres Farinha pelo apoio e contributo para esta Tese.

A todo o corpo docente do Mestrado em Sistemas Avançados de Gestão em Saúde por todos os novos horizontes.

À minha esposa, cuja paciência e empenho nunca permitiu que o desânimo ou o cansaço desencorajassem este percurso. Aos meus filhos, cuja resiliência vai desculpando as faltas na parentalidade. Aos meus pais, avós e irmão, por me terem conduzido à pessoa que sou hoje.

Aos meus colegas do Mestrado, pelo companheirismo, pelos momentos e conhecimentos partilhados.

Aos meus colegas de trabalho, cuja colaboração tornou possível este meu percurso.

Aos meus amigos, aqueles que sempre me apoiaram.

A todos aqueles que participaram voluntariamente, pela sua disponibilidade e atenção.

ÍNDICE

1 Introdução 19

2 Inteligência Artificial na Saúde 22

3 Inteligência Artificial na neonatologia e cuidados intensivos neonatais 26

3.1 Uma perspetiva global 26

3.2 Abordagem baseada em problemas na UCIN – alguns exemplos 28

3.2.1 O Problema da Infeção em Neonatologia 28

3.2.2 Papel da IA na Sépsis Neonatal 30

3.2.3 Monitorização Não-Invasiva 32

3.2.4 IA na Imagiologia Neonatal 34

3.2.5 Ventilação do RN Prematuro e IA 36

3.2.6 Retinopatia da Prematuridade 37

3.2.7 Monitorização da Dor 38

3.3 A IA na Eficiência e Qualidade 40

4 Aplicação de IA na UCIN da ULS 42

4.1 O Problema da Nutrição Adequada no RN 42

4.2 Extração de Elementos Clínicos 44

4.2.1 Natureza do Texto Clínico e Desafios da Extração 44

4.2.2 Abordagens Baseadas em Regras e Padrões 45

4.2.3 Sistemas Baseados em Estatística e Aprendizagem Automática 45

4.2.4 Aprendizagem Profunda e Modelos de Linguagem Contextuais 46

4.2.5 Integração de MedSpaCy e Pipelines Híbridos 47

4.2.6 Large Language Models na Prática Clínica – e porque não? 48

5 Large Language Models 50

5.1 Introdução 50

5.2 Arquitetura Básica dos LLMs 51

5.2.1 Transformer 51

5.2.2 Tipos de Arquitetura: Encoder-only, Decoder-only e Encoder-Decoder 60

5.3 Prompts e Prompt Engineering 62

5.4 Optimização, Fine-Tuning, Ajustes 67

5.4.1 RAG – Retrieval Augmented Generation 67

5.4.2 Fine-Tuning Supervisionado, Instruction Tuning, Fine-Tuning específico de domínio, Multi-task Fine-Tuning e Chain of Thought Supervisionado 69

5.4.3 Multi-modality Fine-Tuning 70

5.4.4 Parameter-Efficient Fine-Tuning 71

5.5 Problemas Inerentes aos LLM 73

5.5.1 Alucinação 73
5.5.2 Viés e Equidade 77
5.5.3 Interpretabilidade 79
5.5.4 Contexto e Memória 79
5.5.5 Segurança 79
5.5.6 Privacidade e Legalidade 80
5.5.7 Outras limitações 80

5.6 Aplicabilidade e Exequibilidade 81

6 Setup Experimental do projeto 82

6.1 Dados – perspetiva CRISP-DM 83

6.2 Apresentação do Projeto da Tese 88
6.2.1 Clinical Understanding (Business Understanding) 88
6.2.2 Data Understanding 90
6.2.3 Data Preparation 94
6.2.4 Modelagem 95
6.2.5 Avaliação 95
6.2.6 Clinical Deployment 95

6.3 Dataset Sintético 96
6.3.1 Metodologia de Geração Sintética dos Textos 96
6.3.2 Abordagem Geral 97
6.3.3 Estratégia de *Prompting* 98
6.3.4 Comparação entre Modelos 101
6.3.5 Controlo e Validação dos Resultados 102
6.3.6 Validação do Dataset e Limitações – Análise Crítica 104
6.3.7 Aplicações 104

7 Modelo LLM a Utilizar 105

7.1 Hardware 105

7.2 Método de Seleção do LLM 106
7.2.1 Fluxo Sistematizado de Extração de Variáveis 106
7.2.2 Métricas Utilizadas para Avaliação dos Modelos 109
7.2.3 Modelos Considerados para Análise 110
7.2.4 Resultados – Extração de Variáveis do Dataset 111
7.2.5 Resultados Globais dos Modelos 116

7.3 Comentários ao Desempenho dos Modelos 117

8 Modelação e Avaliação de Resultados 118
8.1.1 Descrição do Processo do Ponto de Vista Humano 118
8.1.2 Organização do Processo Automatizado com IA 119

8.2 Processo Detalhado 119

8.2.1 Tarefa 1 - Extração da Informação Clínica Relevante 121
8.2.2 Tarefa 2 – Cálculo de Z-Scores 123
8.2.3 Tarefa 3 - Identificação das Referências Nutricionais adequadas ao RN 125
8.2.4 Tarefa 4 - Comparação com os Standards Adequados e Identificação de Ajustes para Correção 127
8.2.5 Tarefa 5 – Conclusão e Elaboração de uma Proposta Estruturada 129
8.3 Avaliação dos Resultados 133
9 Discussão 136
9.1 Contribuições 136
9.2 Resposta às Questões da Investigação 137
9.3 Implementação futura em UCIN 138
9.4 Perspetivas Futuras 139
10 Conclusão 140
11 Referências Bibliográficas 142

ÍNDICE DE FIGURAS

Figura 1 - Especialidades médicas e estudos de IA [1] 23

Figura 2 – publicações e citações sobre IA e Cuidados Intensivos [3]. 23

Figura 3 - Modelos Básicos de IA [1]. 24

Figura 4 - Exemplo de imagem Power Spectral Density. Neste caso, a variável analisada foi a variabilidade da frequência cardíaca [23]. 32

Figura 5 – Comparação entre a FC (elétrodos) e a FC calculada [26] 33

Figura 6 – RN prematuro em fototerapia (luz azul – 460nm) 33

Figura 7 – RN PT, 24 semanas, posicionado com material de conforto [28] 34

Figura 8 – Identificação de sonda nasogástrica (NGT), tubo endotraqueal (TET), cateter venoso umbilical (UVC) e cateter arterial umbilical (UAC) [38] 35

Figura 9 – Predição do volume minuto em ventilação AC-VG com base nos 5 ciclos anteriores. Volume minuto é o volume total de gás que entra (ou sai) do pulmão por minuto. É igual ao Volume Corrente (VC) multiplicado pela frequência respiratória (f): MV = VC x f [41] 36

Figura 10 – Diagnóstico remoto de ROP com *deep learning* e *cloud computing*. Neste sistema, apenas os dados pré-processados das imagens da RetCam© são enviados para uma *cloud* [44]. 37

Figura 11 – Modelo proposto [47] 39

Figura 12 – Flowchart Classifier Fusion para predição de Length of Stay [51]. 40

Figura 13 – Evolução do Aporte Nutricional em Volume ao longo da primeira semana de vida 43

Figura 14 – Comparação na performance de extração de dados clínicos de múltiplos modelos baseados em LLM com a referência BERT (RoBERTa SQuAD, já previamente treinado para o efeito) [56] 48

Figura 15 – LLMs na medicina em geral [70] 49

Figura 16 – Esquema Transformer [73] 51

Figura 17 – Arquitetura global Encoder-Decoder [73] 52

Figura 18 – Idealização da representação de palavras após tokenização e embedding: representação num espaço bidimensional – palavras semelhantes "ocupam" localizações próximas [75]. 53

Figura 19 – Positional Encoding [73] 54

Figura 20 – Mecanismo de Atenção [73] 56

Figura 21 – Output final [73]. 56
Figura 22 – Esquema global [73]. 57
Figura 23 – Esquema final completo da arquitetura Codificador → Descodificador *(Encoder – Decoder)* [73] 59
Figura 24 – Processo de amostragem Top-P e Top-K [82]. 63
Figura 25 – Tree of Thoughts (ToT) [81]. 65
Figura 26 – Comparação entre o LLM como "Agente" (ReAct prompting) e Agentic AI [84]. 65
Figura 27 – Visão global do processo RAG [85]. 68
Figura 28 – Processo exemplificativo de *Instruction Tuning* [86]. 69
Figura 29 – Arquitetura de um MLLM [87] 71
Figura 30 – LLMs: tipos de viés na prática clínica [93] 78
Figura 31 - Esquema clássico da metodologia CRISP-DM [100]. 83
Figura 32 – Fase 1 com as adaptações propostas [100]. 84
Figura 33 – Fase 2 com as adaptações propostas [100]. 85
Figura 34 – Fase 3 com as adaptações propostas [100]. 86
Figura 35 – Fase 4 de acordo com CRISP-MED-DM com as adaptações propostas [100]. 87
Figura 36 – Fase 5 com as adaptações propostas [100]. 87
Figura 37 - Exemplo genérico de texto a utilizar (NB – dados fictícios) 91
Figura 38 – Violin Plot: a) distribuição de idade gestacional por sexo (note-se a distribuição com maior frequência em torno das 27s e depois novamente, mas em menor número em torno das 32s e superiores); b) raciocínio idêntico, mas com o peso de nascimento. 92
Figura 39 – Distribuição por sexo de a) Aporte Hídrico Total, b) Calorias Totais, c) Proteínas e d) Lípidos. 93
Figura 40 – Distribuição por sexo e idade gestacional: a) Aporte Hídrico Total (ml/kg/dia) e b) Aporte Calórico Total (kcal/kg/dia) 93
Figura 41 – Esquema global do processo 96
Figura 42- Exemplo genérico de texto a utilizar (dados fictícios) 97
Figura 43 – Exemplo de texto clínico sintético, considerado não conforme pela falta de complexidade e simplificação absurda. Deve ser referido que o parágrafo com a linha sobre aportes foi respeitada e encontra-se na posição habitual. Podemos ainda encontrar alguns exemplos de pormenores atribuíveis a um modelo generalista / não específico de medicina, como a notação DBP leve (classificação inexata) ou ainda a

omissão de itens (Renal e Infecioso, por exemplo). Texto gerado com o DeepSeek©. 98

Figura 44 – Diário Clínico Sintético, considerado de boa qualidade. GPT 5©..... 99

Figura 45 – Prompt utilizada em todos os modelos. Tal como explicitado ao modelo, foi fornecido um glossário em arquivo .txt adicional ("notas ao modelo.txt"), para esclarecer dúvidas de siglas e acrónimos comuns nos diários clínicos. Foram também fornecidos os exemplos originais. 99

Figura 46 – "Notas ao modelo.txt" – pequeno glossário fornecido ao modelo para esclarecer as siglas e acrónimos utilizados. 100

Figura 47 – Fluxo da depuração do Dataset Sintético 103

Figura 48 – Esquema de funcionamento para extração de variáveis a partir de texto clínico 107

Figura 49 - Comparação do tempo total para processar os 142 textos clínicos e extrair as variáveis pedidas. **a)** apresentam-se as relações entre tempo total (min; eixo y), tamanho (Gb; eixo x) e o número de parâmetros (Dimensão parâmetros em milhares de milhão; cores). **b)** Tempo total de processamento para comparação. 111

Figura 50 – Comparação entre os modelos: a) MAE; b) RMSE; c) Missed Rate %; d) MAPE. 115

Figura 51 – Processo Simplificado de Ajuste de Aportes Nutricionais 118

Figura 52 – Diagrama do Processo Global 120

Figura 53 – Diagrama da Tarefa 1 121

Figura 54 – **Tarefa 1**: a) carregar "lista de variáveis.txt" contendo as variáveis a extrair e o texto clínico a analisar "RN ##.txt"; b) conversão do texto simples da lista de variáveis para chave canónica (gera o JSON #1 vazio); c) LMStudio corre o LLM com os parâmetros ajustados; d) variáveis extraídas e vertidas em formato normalizado no JSON #2. 122

Figura 55 – Diagrama da Tarefa 2: 1-2) o referencial de Z-Scores (-3 a 3 para cada variável somatométrica, idade gestacional e sexo) em .csv é carregado através do interface do Gradio; 3-4) o JSON #2 com os dados extraídos do texto clínico é utilizado para calcular o Z-Score para cada variável somatométrica (valores obtidos por interpolação com base no referencial em .csv); 5-6) os Z-Scores calculados são vertidos no JSON #3 para utilização posterior. 123

Figura 56 – **Tarefa 2**: a) carregar referências "Z-Scores.csv"; b) tabela com Z-Scores calculados para as variáveis extraídas (dados contidos no JSON #2); c) JSON #3 com os Z-Scores calculados. 124

Figura 57 – Diagrama da Tarefa 3 125

Figura 58 – **Tarefa 3**: a) carregar referências para aportes nutricionais "aportes.csv"; b) tabela com os intervalos de aportes estandardizados para as características do RN; c) JSON #4 com os intervalos admissíveis para o RN em análise. 126

Figura 59 – Diagrama da Tarefa 4 127

Figura 60 – **Tarefa 4**: 128

Figura 61 – Prompt final para gerar uma conclusão e proposta de ajuste em linguagem natural 130

Figura 62 – Diagrama da Tarefa 5 130

Figura 63 – **Tarefa 5**: 131

Figura 64 – Exemplo genérico de texto a utilizar (NB – dados fictícios) 161

Figura 65 – Fluxo de trabalho esquematizado 162

Figura 66 – Prompt inserida 163

Figura 67 – Interface "gráfico" no LMStudio. Permite ajustar temperatura, top.p e top.k 163

Figura 68 – Interface "gráfico" e Neonatologista-friendly para manipular os textos de diário clínico e, na versão final, permite o download das tabelas .csv 164

Figura 69 – Tabela csv com os 10 textos sintéticos de diário. Ainda algumas falhas. 165

Figura 70 – Utilização de textos não médicos (*decoy)* 165

Figura 71 – Tabela csv e tabela com time-stamp geradas pela versão final 166

ÍNDICE DE TABELAS

Tabela 1 - Lista de Variáveis, com mapeamento LOINC e SNOMED-CT para definir terminologia e semântica. 89

Tabela 2 – Estatística Descritiva do Dataset 92

Tabela 3 – Comparação Qualitativa dos diferentes LLMs utilizados 101

Tabela 4 – Métricas de Classificação 109

Tabela 5 – Métricas de Erro 109

Tabela 6 – Modelos testados (parâmetros, tipo de quantização, tamanho e memória "real" ocupada com o funcionamento de todas as aplicações); B = milhares de milhão de parâmetros 110

Tabela 7 – Tempo de processamento e características dos modelos 111

Tabela 8 – Scores para variáveis categóricas. 113

Tabela 9 a) e b) – Scores das variáveis categóricas. Em a), calculados para classe M e em b), calculados para classe F. Em cada classe, os valores em falta foram contabilizados como FN (falso negativo). 114

Tabela 10 – MAE, RMSE, MAPE, Missed e Missed Rate (%) dos modelos testados. 115

Tabela 11 – Resultados obtidos na amostra com 30 textos clínicos provenientes do dataset 134

LISTA DE SIGLAS, ACRÓNIMOS E ABREVIATURAS

Sigla	Significado
APACHE II	Acute Physiology and Chronic Health Evaluation, version II
CNN	Redes Neuronais Convolucionais
CPAPn	Continuous Positive Airway Pressure - nasal
CRISP-DM	Cross Industry Standard Process for Data Mining
CRISP-MED-DM	Cross Industry Standard Process for Data Mining in Medicine
FC	Frequência Cardíaca (em batimentos por minuto)
FR	Frequência Respiratória (em ciclos por minuto)
GAN	Redes Neuronais Adversariais Generativas
IA	Inteligência Artificial
IG	Idade Gestacional (em semanas e dias)
IPM	Idade Pós-Menstrual (em semanas e dias)
LLM	Large Language Model
LOINC	Logical Observation Identifiers Names and Codes
MoE	Mixture of Experts
ML	Machine Learning
NER	Named Entity Recognition
NIH	National Institutes of Health
NLP	Natural Language Processing
PCR	Proteína C-Reactiva
PCT	Procalcitonina
RAG	Retrieval Augmented Generation
ROP	Retinopatia da Prematuridade
RN	Recém-Nascido
RNN	Redes Neuronais Recorrentes
RNMBP	Recém-Nascido de Muito Baixo Peso (<1500 g)
SNT	Sépsis Neonatal Tardia
SNOMED-CT	Systematized Nomenclature of Medicine - Clinical Terms
SpO2	Saturação Periférica de Oxigénio (em %)
UCIN	Unidade de Cuidados Intensivos Neonatais
VON	Vermont Oxford Network

1 INTRODUÇÃO

No decurso desta Tese, investiga-se a aplicabilidade da Inteligência Artificial (IA) em contexto de Unidade de Cuidados Intensivos (UCIN).

Apesar de serem múltiplos os exemplos de utilização, sobretudo em fases exploratórias e sem grandes soluções já comercializadas, a sua pertinência é cada vez mais incontornável. No seu conjunto, têm sobretudo explorado opções baseadas em ML e interpretação de grandes volumes de dados, geralmente provenientes de vários monitores de sinais vitais e semelhantes; com a intenção de prever eventos *major*, como sépsis, ou até melhorar parâmetros de ventilação. Outros têm procurado soluções baseadas em CNN (Redes Neuronais Convolucionais) para interpretação mais ou menos autónoma de imagens, com maior ou menor sucesso.

A utilização de modelos baseados em Natural Language Processing (NLP) e Large Language Models (LLM) apostam na facilidade de interação com a equipa clínica e, seja através de Retrieval Augmented Generation (RAG) extenso, seja através de Mixture of Experts (MoE) altamente específicos, apresenta soluções comercializadas para ajuda na tomada de decisão (ex. ClinicalKey© da Elsevier, entre outros). Também frequentemente somos tentados a usar soluções mais acessíveis, como o ChatGPT©, Gemini©, ou outros para auxiliar em pequenas questões.

Como pontos fortes, podemos indicar que na sua maioria são soluções robustas, algumas muito mais específicas e adaptadas para o domínio médico (mas não são gratuitas ou de livre acesso) e, na sua maioria, correspondem a verdadeiros modelos *state of the art*. Como pontos negativos, não podemos ignorar a necessidade de obrigarem, de certo modo, a utilizar alguns dados, muito dependentes da observação de boas práticas pela equipa clínica; e o facto de se localizarem em "parte incerta". Estas questões colocam muitos entraves na utilização de todo o seu potencial.

A questão fundamental seria se podemos encontrar uma solução de compromisso: por um lado, respeitando a legislação sobre utilização de dados clínicos pessoais e por outro, explorando algum do potencial dos LLM.

Então, naturalmente, surgiu uma proposta de trabalho dedicada à neonatologia e cuidados intensivos neonatais. Neste campo, apesar de vasto, existem sempre aspetos mais modestos que são esquecidos na intensidade do trabalho diário.

Um desses aspetos é a nutrição adequada dos RN: esta deve fornecer todos os nutrientes necessários a um ser em crescimento (deve tentar replicar o ambiente intrauterino durante a gravidez); deve fazer face a uma adaptação "forçada" a um novo ambiente para o qual transitou demasiado cedo (a prematuridade e todas as

agressões daí resultantes) e ainda comportar todas as intercorrências desfavoráveis (como a sépsis).

No âmbito desta Tese, o papel do LLM como assistente de atividade clínica foi testado num cenário específico:

"Qual a capacidade de um LLM para elaborar uma proposta de ajuste nutricional com base no diário clínico do RN em UCIN?"

Assim, o Objetivo da Tese é:

"Testar a capacidade de um LLM para elaborar uma proposta de ajuste nutricional individualizada, partindo do texto do diário clínico, num processo limitado pelo hardware habitual na ULS e excluindo qualquer processamento online."

Esta questão é, em si mesma, inovadora e disruptiva, ao tentar replicar uma estratégia mínima "Agentic AI" e testar os limites do LLM neste cenário. Esta abordagem, completa, não foi ainda explorada em publicações científicas.

Neste percurso, pioneiro, foram encontrados e ultrapassados vários desafios:

1. idealização de todo o processo completo, desde a extração de dados até à geração de uma proposta, sempre com o LLM na posição central;
2. exploração dos ajustes possíveis no LLM para a função específica;
3. necessidade de montar um processo de extração e registo de dados clínicos relevantes, a partir de um texto clínico não estruturado;
4. estratégias para minimizar problemas relativos ao funcionamento probabilístico dos LLM;
5. metodologia de avaliação do processo de extração;
6. metodologia de avaliação de diversos LLM para escolha do mais ajustado;
7. necessidade de criar e validar um dataset sintético, apesar do pedido para a Comissão de Ética da ULS autorizar a utilização de dados reais ainda não ter tido um parecer;
8. idealização do processo de consulta e comparação com os standards;
9. avaliação do resultado final, ou seja, a proposta gerada pelo LLM;
10. e, finalmente, o imperativo de utilizar soluções simples e compatíveis com os equipamentos mais acessíveis, pois tudo teria de funcionar apenas na rede da ULS.

Para dar resposta a estes desafios, foi necessário utilizar várias estratégias, das quais destacamos:

a. elaboração de *prompt* e utilização de *prompt engineering* (na extração de dados; na elaboração da proposta; na criação do dataset sintético);
b. avaliação objetiva do desempenho dos vários LLM com métricas reconhecidas para classificação e erro;
c. organização do processo de acordo com a metodologia CRISP-MED-DM;
d. exploração de estratégias para partilha de dados relevantes entre as várias etapas do processo.

Assim, esta Tese está estruturada do seguinte modo:

- Uma breve revisão de soluções de IA na Saúde e na Neonatologia, onde abordam soluções *state of the art* – capítulos 2 e 3.
- O enquadramento do problema da nutrição no RN e concretização da questão de investigação, abordando algumas estratégias possíveis – capítulo 4.
- No capítulo 5 é explorada a temática dos LLM, abrangendo o seu funcionamento e estrutura, para que se possa entender as suas limitações e problemas; e também delinear estratégias para as minimizar.
- O processo de extração de dados é apresentado no capítulo 6, de acordo com a metodologia CRISP-MED-DM. Neste capítulo também é descrito todo o processo de elaboração do dataset sintético utilizado.
- A escolha do LLM a utilizar, com todo o processo de avaliação e resultados das métricas, estão amplamente descritos no capítulo 7.
- O capítulo 8 é dedicado à Modelação e Avaliação de Resultados. Ao longo do mesmo descreve-se, pormenorizadamente, todo o processo deste Projeto: desde a extração de dados, processamento, comparação com standards, classificação e, por fim, a elaboração de uma proposta de ajuste em linguagem natural. No final, encontra-se a avaliação dos resultados.
- No capítulo 9 é apresentada a Discussão do Projeto, com as contribuições que este adiciona na área da IA mas também na Medicina. São ainda referidas algumas possibilidades de expansão futura, nomeadamente a implementação em ambiente real na UCIN e novas abordagens baseadas no processo desenvolvido.
- Por fim, a Conclusão Final da Tese apresenta a súmula de todos estes pontos no capítulo 10.
- Nos Anexos encontram-se: 1) artigo publicado contendo o *proof of concept* deste Projeto, apresentado no Congresso Internacional PAMDAS 2025; 2) descrição do processo inicial exploratório; 3) Exemplos de textos clínicos (do dataset sintético); 4) Tabelas csv elaboradas com os standard nutricionais e Z-Scores.

2 INTELIGÊNCIA ARTIFICIAL NA SAÚDE

Inteligência Artificial na Medicina

Inteligência Artificial pode ser definida como um conjunto de algoritmos que replicam a função cognitiva humana, utilizando modelos produzidos a partir da análise estatística de grandes volumes de dados [1]. Constitui uma área de pesquisa na informática dedicada ao desenvolvimento de métodos e software que permitem a perceção do ambiente e a interação com ele através da aprendizagem, culminando na realização de um objetivo ou tarefa. Esses sistemas já estão integrados na rotina diária, e podem ser exemplificados por, mas não se limitando a: Google (motor de busca), YouTube (sistema de recomendação e pesquisa), Alexa e Siri (assistentes virtuais); ChatGPT, DeepSeek, Gemini, entre outros.

A IA constitui uma disciplina académica com 70 anos de evolução[2], com uma evolução marcada por ciclos de progresso mais rápido entrecortados por épocas de estagnação (os invernos da IA), muito dependentes de investimento económico. A evolução exponencial dos últimos 5 anos, especialmente com o aparecimento da IA generativa, levantou preocupações com a segurança e interferência na vida humana e a necessidade de impor regulamentação.

Este crescimento exponencial tem vindo a retirar a IA do laboratório e a integrá-la no dia a dia, quer na vertente pessoal, quer na vertente profissional.

A sua aplicação na medicina tem sido crescente, em especial na medicina de adultos (ver Figura 1): praticamente 50% corresponde a áreas médicas (patologia clínica lidera), sendo os restantes 50% divididos por área cirúrgica, imagem médica e psiquiatria, sendo que a área pediátrica corresponderia a menos de 10% do total [1].

Figura 1 - Especialidades médicas e estudos de IA [1]

Nos últimos 20 anos, o número de referências sobre IA e cuidados intensivos teve um crescimento exponencial, o mesmo se verifica nas pesquisas e citações (ver Figura 2) [3], [4] .

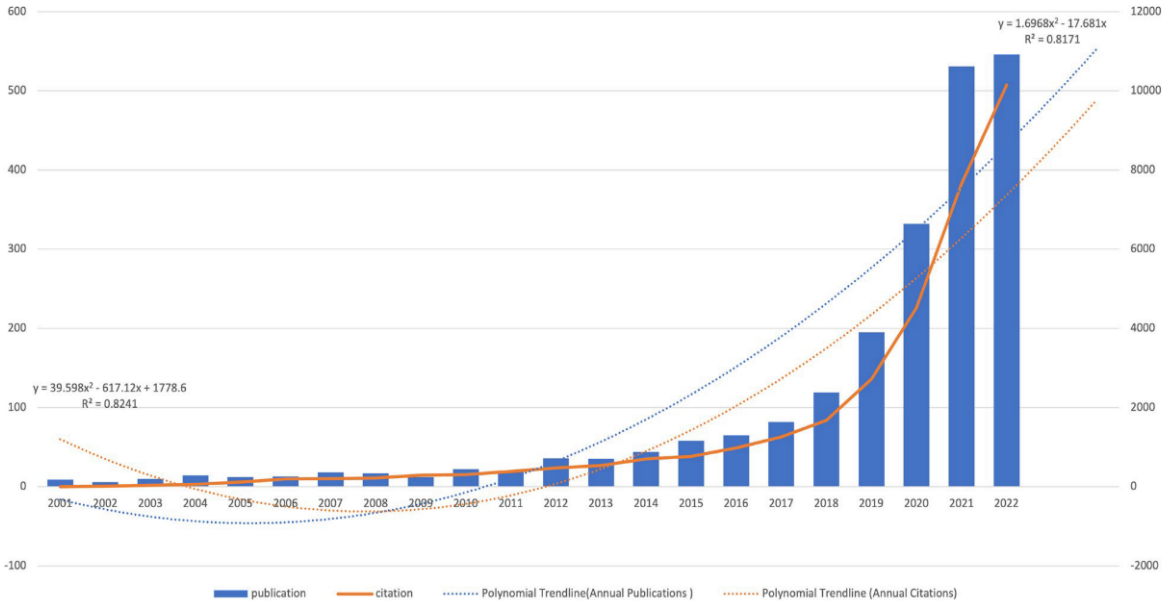


Figura 2 – publicações e citações sobre IA e Cuidados Intensivos [3].

O constante aperfeiçoamento e melhoria nos modelos tem permitido aplicações cada vez mais ajustadas e úteis na prática clínica. Os modelos baseados em machine learning (ML) partem da análise de grandes volumes de dados para gerar um algoritmo que produz um resultado, habitualmente não imediatamente óbvio através da análise clássica. Na prática clínica, ajudam na tomada de decisões ao facilitarem a integração de múltiplas informações [1], [5]. Como exemplo, temos alguns calculadores de risco, como o APACHE II. Este permite um cálculo preditivo para a estratificação de pacientes com doença severa em unidades de cuidados intensivos e utilizou os registos de 5815 admissões em 13 hospitais, conjugando 12 parâmetros fisiológicos, para além da idade e estado de saúde prévio. O resultado do score, para além de gerar um prognóstico, permite ainda a comparação da performance de diferentes unidades para patologia semelhante. É uma referência desde 1985 [6].

Genericamente, um processo baseado em ML acabará sempre por produzir resultado semelhantes perante situações similares, visto o algoritmo ter sido aperfeiçoado para isso mesmo, após análise de grandes quantidades de dados. Mas este processo tem as suas limitações: necessita de dados de qualidade, muitas vezes previamente trabalhados, e acabam por, "no limite", serem processos estatísticos altamente complexos.

O passo lógico seguinte seria a tentativa de replicar o padrão de processamento cerebral, ao tentar recriar uma rede neuronal.

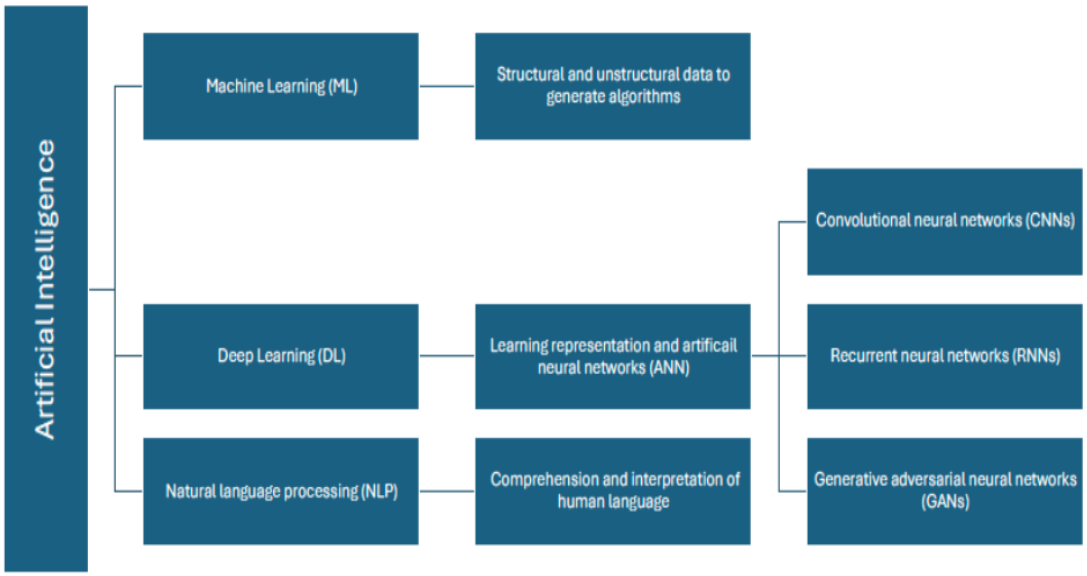


Figura 3 - Modelos Básicos de IA [1].

A replicação do padrão de processamento cerebral passa, para já, pela criação de redes multicamadas, que podem ser genericamente divididas em redes neuronais convolucionadas (CNN), redes neuronais recorrentes (RNN) e redes neuronais adversariais generativas (GAN). A relação entre estes modelos encontra-se na Figura 3.

Resumidamente, as CNN são utilizadas para analisar imagens e vídeos, sobretudo dados espaciais, habitualmente subdivididos numa grelha. Assim, a sua aplicação mais habitual é no reconhecimento facial, reconhecimento de objetos e na imagem médica / diagnóstico por imagem. As GAN colocam duas CNN a trabalhar em sequência, em que uma gera dados e a outra analisa-os. Desta forma, através da interação entre ambas, estão continuamente a melhorar-se reciprocamente. São habitualmente utilizadas para criar imagens / vídeo, mas podem gerar imagens impossíveis de existir. Infelizmente estão a ser utilizadas para criar *deepfakes*... As RNN analisam séries temporais longas, especificamente a relação, a sequência e a ordem dentro dessa série, sendo utilizadas na análise de texto, discurso, som, entre outros, desdobrando as relações complexas com base em regras de semântica e sintaxe, mas têm sido substituídas por NLP (Natural Language Processing) evoluídos – os Large Language Models (LLMs). Estes são muito mais poderosos e permitem ainda interação e geração de texto em linguagem natural [5], [7].

Nos últimos anos temos assistido a uma verdadeira revolução nos cuidados de saúde com a utilização dos LLM como ferramenta ou como assistente médico, ao permitir gerar respostas a questões relacionadas com diagnóstico, terapêutica, seguimento, etc.

3 INTELIGÊNCIA ARTIFICIAL NA NEONATOLOGIA E CUIDADOS INTENSIVOS NEONATAIS

3.1 Uma perspetiva global

Na Neonatologia e Cuidados Intensivos Neonatais estamos a viver uma mudança de paradigma: passamos de calculadores de risco para a possibilidade de análise de tendências em tempo real e até à integração da IA como parceiro na atividade clínica.

As primeiras referências sobre a aplicação de IA em Cuidados Intensivos Neonatais remontam à década de 1990-2000 com a monitorização neurológica, como a deteção automática de eventos em eletroencefalograma. Este trabalho resultou da colaboração de 3 centros de referência, com 55 recém-nascidos e utilizou cerca de 280 horas de registos de electroencefalogramas, para que se obtivesse uma taxa de deteção de 70%. Este método, resumidamente, comparava a ocorrência de picos de atividade elétrica com o padrão de base anterior, permitindo a identificação de crises [8].

A Neonatologia e as Unidades de Cuidados Intensivos Neonatais (UCIN) constituem um bom campo experimental para aplicar a IA na prática clínica. Numa UCIN estão internados RN com patologia, habitualmente durante longos períodos (são frequentes internamentos de 90 dias no caso de prematuros de 24 semanas de idade gestacional). Durante este período, são produzidas grandes quantidades de dados, que compreendem a monitorização contínua multi-parâmetro dos monitores de sinais vitais, avaliações não automatizadas de outras variáveis fisiológicas como a temperatura corporal, variáveis somatométricas e sua relação com percentis / Z-scores (peso, comprimento, perímetro cefálico, etc), escala de dor, tolerância alimentar, diurese, atividade e reatividade espontânea. Todos estes dados são gerados num espaço físico (UCIN) relativamente contido e controlado, de dimensões "relativamente" reduzidas quando comparado com o universo do doente adulto [5], [7].

Podemos referir variados exemplos de aplicações:

- A mais importante é sem dúvida a Vermont Oxford Network, que se estende muito para além desses dois locais, englobando já cerca de 1300 UCINs internacionais. Existe desde 1988 e é uma organização sem fins lucrativos. Teve um papel fundamental na implementação de várias medidas nas unidades, como a uniformização de práticas ventilatórias, restrição de uso de oxigénio, primazia do leite materno na alimentação do prematuro, mas sobretudo pela utilização de benchmarking, levou à comparação de várias estratégias entre as UCINs com ganhos claros na qualidade. E, claro, disponibiliza um calculador de risco e de mortalidade estimada de acordo com alguns parâmetros como idade gestacional, peso, entre outros. Um

semelhante e de utilização pública está disponível através do portal do National Institutes of Health[1].

- Monitorização não invasiva de sinais vitais com recurso a câmaras e processamento das imagens por IA;
- Análise e interpretação de imagens com IA;
- Auxílio na ventilação invasiva neonatal com análise de parâmetros baseada em IA;
- Deteção automatizada de retinopatia da prematuridade através de análise de imagens de fundoscopia;
- Auxílio na terapêutica da dor em UCIN, através da análise de expressões faciais.

[1] Acessível em: https://www.nichd.nih.gov/research/supported/EPBO/use (último acesso em 5 de dezembro de 2025).

3.2 Abordagem baseada em problemas na UCIN – alguns exemplos

Em Neonatologia e Cuidados Intensivos Neonatais podemos afirmar que todos os problemas de uma unidade de cuidados intensivos de adultos se mantêm, mas com pelo menos três dificuldades acrescidas:

1) a incapacidade do RN se queixar e permitir o ajuste de algumas atitudes clínicas, restando apenas uma bateria de análises e exames, e o senso clínico;
2) a fragilidade inerente a um ser em crescimento e maturação, pois todos os órgãos se encontram ainda não totalmente funcionais:
   a) condiciona várias alterações na normal fisiologia que devem ser compensadas, corrigidas e levadas em conta com a adaptação de doses de fármacos, por exemplo;
   b) esses mesmos órgãos continuarão o seu processo de maturação e crescimento, mas ao contrário do ambiente uterino em que estão "apenas" em crescimento, agora terão ainda de manter a sua função, quase sempre em esforço e de modo deficitário;
   c) e, para além de todas as agressões a que estão submetidos durante este período, a sua maturação ficará quase sempre modificada não apenas pelo ambiente diferente, mas também pelas "cicatrizes" que ficam;
3) o tamanho dos RN prematuros – são cada vez mais frequentes pesos inferiores a 1000g e mesmo abaixo das 600g. Isto coloca problemas no material utilizado, nos limites da própria monitorização, quantidade de parâmetros monitorizáveis em simultâneo e na falta de soluções adequadas.

3.2.1 O Problema da Infeção em Neonatologia

A Sepsis Neonatal Tardia (SNT) é definida pelo isolamento de uma bactéria patogénica em cultura de sangue (hemocultura) ou líquido cefalorraquidiano (LCR), ou fungo no sangue, obtidas após as 72 horas de vida [9], [10] Não deve ser confundida com a Sépsis Neonatal Precoce, que define de modo idêntico, exceto que ocorre até às primeiras 72 horas de vida do RN. Estão implicados fatores de risco e mecanismos diferentes, por isso são entidades diferentes, com abordagens clínicas distintas.

A ocorrência de um episódio de SNT traduz-se uma agressão ao RN, sendo mais frequente e mais severa quanto maior a fragilidade deste. Assim, este problema afeta sobretudo RN prematuros, com maior incidência em prematuridades mais extremas e com maior severidade nos *outcomes* [9], [10].

A importância desta entidade em cuidados intensivos neonatais relaciona-se não apenas pela mortalidade inerente a cada episódio de SNT, variando entre 4-15%, mas também pelo seu impacto negativo no neurodesenvolvimento [9], [10], [11]. A lesão do Sistema Nervoso Central (SNC) resulta da citotoxicidade bacteriana direta e da inflamação sistémica adversa. Esta pode ocorrer mesmo sem invasão do agente patogénico no SNC, bem como por alteração da perfusão cerebral, no contexto de instabilidade hemodinâmica. De uma maneira geral, estes recém-nascidos de muito baixo peso (RNMBP) apresentam um risco até 10 vezes superior de alterações neurocognitivas aos 5 anos. Para este risco contribuem não apenas as lesões diretas nos casos de meningite, mas também as alterações na substância branca, condicionando leucomalácia, que consoante a localização, assim se traduzirá em quadros clínicos com alterações sensitivo-motoras (paralisia cerebral) e/ou em alterações cognitivas diversas, condicionantes na performance escolar futura.

Apesar dos esforços, em Portugal, ao analisarmos os dados do registo nacional do muito baixo peso nos últimos 15 anos, constatamos que num total de 15500 de RNMBP, cerca de 20% teve, pelo menos, um episódio de sépsis tardia (não publicado). E, infelizmente, este valor tem-se mantido estável ao longo dos anos. Estes valores são superiores aos referidos noutras séries, em redes de maiores dimensões, como a Vermont-Oxford Network, onde são referidos valores de 9 a 14% [10], [11]. Assim, há uma necessidade de intervir para se conseguirem valores mais próximos dos referidos, quer na capacidade de reduzir os falsos positivos, quer no tempo até ao diagnóstico.

Na literatura encontramos diversos trabalhos que referem a necessidade de aplicar a IA no diagnóstico da SNT. Destes, muitos referem trabalhos em curso e outros publicados apresentam já resultados [1], [5], [7], [9], [10], [12], [13], [14]. A evidência relatada coloca, segundo o meu ponto de vista, a IA como fazendo parte da prática clínica, como um meio complementar de diagnóstico e terapêutica ou ainda como uma vulgar análise de proteína C-reativa ou procalcitonina. Com efeito, a IA consegue integrar múltiplos conjuntos de parâmetros fisiológicos, analisar tendências e gerar um pré-alerta. Se a isto juntarmos alguns resultados de análises, devidamente processados por um algoritmo de ML, podemos não apenas ter diagnósticos mais rápidos, mas também reduzir eficazmente os falsos positivos. A redução de falsos positivos permitiria reduzir a pressão antibiótica, reduzir resistência antibiótica futura e tornar a gestão antibiótica mais simples. Na literatura é referido que até 31% da mortalidade por sépsis pode ser atribuída a resistência antibiótica [15].

3.2.2 Papel da IA na Sépsis Neonatal

A IA, neste exemplo, assume uma importância que se projeta muito para além do benefício do RNMBP enquanto indivíduo (fundamental) mas também em conceitos mais abrangentes de controlo de infeção hospitalar e medidas de saúde pública.

Na análise de soluções e artigos semelhantes encontramos uma abordagem comum: praticamente todos utilizaram parâmetros fisiológicos cuja variação e tendência se alteram no início e durante a SNT [15], [16], [17], [18], [19], [20]. A influência da resposta inflamatória sistémica na resposta do sistema nervoso autónomo está amplamente descrita, podendo condicionar a variabilidade em parâmetros fisiológicos hemodinâmicos, respiratórios e autoregulação térmica.

De entre os vários modelos, o denominador comum entre todos os parâmetros é a Frequência Cardíaca (FC) - variabilidade e tendência. Este parâmetro já é utilizado numa aplicação e aparelho comercializados – HeRO© Heart Rate Observation System®, Inspiration Healthcare Group. Ainda assim, apesar de ter demonstrado boa sensibilidade, na prática clínica verificou-se que apesar do aumento em 10% do número de hemoculturas e uma maior duração (5%) na antibioterapia, conseguiu uma redução da mortalidade em 22% [21]. Assim, torna-se evidente que devem ser adicionados mais parâmetros. Esta abordagem já foi tentada com sucesso utilizando outro parâmetro fisiológico – a Saturação Periférica de Oxigénio SpO_2 [20]. Através da integração de episódios de apneia, dessaturação, bradicardia, conseguiram resultado superiores aos da FC isoladamente. A Frequência Respiratória (FR) foi derivada do sinal de impedância torácica. A FC e a SpO2 do oxímetro de pulso foram recolhidos de 2 em 2 segundos. A média, o desvio padrão e a correlação da frequência cardíaca, da frequência respiratória e da saturação de oxigénio foram calculadas em janelas de 10 minutos e, posteriormente, foram calculadas as médias para cada hora. A área sob a curva (AUC) neste subconjunto de dados para o modelo de sinais vitais e o índice da FC, isoladamente, foi de 0,684 e 0,707, respetivamente. A combinação num modelo de três variáveis com o índice da FC aumentou a Área Sob a Curva (AUC) em 0,021 para 0,728 (intervalo de confiança (IC) de 95%: 0,010, 0,047; qui-quadrado de Wald = 22,6; P = 0,00001) [20]. Apesar do resultado favorável do algoritmo, os autores ressalvam que não ficou definido um valor a partir do qual deveria ser gerado um alarme. Como fatores menos favoráveis podemos referir a necessidade de pré-processamento dos dados antes da sua inserção no modelo ML [15], [16].

Poderemos inferir do raciocínio anterior que a conjugação de vários parâmetros fisiológicos hemodinâmicos, respiratórios, autoregulação térmica, tolerância alimentar, atividade espontânea será, à partida, mais fiável, específica e com maior poder para alertar na SNT? A análise da literatura não permite tirar conclusões. Não se encontram referências descritas nem com resultados favoráveis nem com maus resultados. A única conclusão possível será a de que ainda não foi tentado.

A segunda questão a colocar aborda a frequência de amostragem dos vários parâmetros fisiológicos: a FC, FR, SpO_2 são avaliados continuamente (limitados pela

taxa de amostragem de cada monitor), mas a TA, a temperatura, a atividade espontânea e a tolerância alimentar são frequentemente apenas avaliados com intervalos de 3 horas. Assim, cada parâmetro teria taxas de amostragem diferentes. Também não é fácil definir um modelo que consiga integrar todos estes dados em tempo real. A maioria das referências trabalha com dados a posteriori. Ao pretendermos adicionar dados analíticos (PCR, PCT, hemograma, entre outros) para melhorar a resposta do algoritmo, encontramos mais um obstáculo: estes não estão disponíveis de modo imediato e implicam a manipulação dolorosa do RNMBP.

A terceira questão diz respeito à dimensão da amostra a considerar. Neste ponto, a revisão da literatura não é muito clara. Alguns autores apresentam modelos funcionais de IA com algoritmos desenvolvidos a partir de dados de 8000 RNMBP. [14]. Outros são menos específicos e referem-se apenas aos dados das próprias unidades, com números mais modestos (n=60), prospetivos e emparelhados caso-controlo, mas ainda assim com bons resultados [15]. Assim, a amostra não parece ser, por si só, uma limitação inultrapassável. E é lícito considerar que a manutenção do modelo preditivo em funcionamento constante permitirá, após a fase inicial de aprendizagem, aperfeiçoar o algoritmo para melhorar a sensibilidade e especificidade.

Resta assim a escolha da melhor abordagem ao algoritmo. A revisão da literatura apresenta resultados muito variados e, apesar de todos optarem por ML, o tipo de processamento oscila entre regressão logística, rede neuronal simples ou CNN.

Na revisão da literatura encontramos algumas soluções diferenciadas com abordagem diferente e que permitem resolver várias das questões enunciadas: integração dos dados diretamente da origem (monitor cardio-respiratório) com sinal direto, não invasivo, em tempo real e ainda utilizando multiparâmetros como a FC e FR. Estes dados são processados e transformados em densidade espectral bi-dimensional e esta é depois analisada no modelo CNN [15]. No exemplo em questão, a amostra utilizada para programação é de apenas 60: 30 casos de sépsis e 30 casos sem sépsis, devidamente emparelhados por idade gestacional e peso. É ainda referido que cada RNMBP contribuiu, por si só, com uma grande quantidade de dados (i.e. foi utilizada a totalidade do tempo de internamento, e ainda no mesmo RNMBP, o período sem sépsis – "normal" e as alterações do mesmo nas horas que antecedem a sépsis). Provavelmente, esta seria a abordagem mais racional. A opção de transformar dados em imagem e depois processar numa CNN é disruptiva (Figura 23), mas a CNN já demonstrou ser capaz de ir muito além (inclusive distinguir fundos oculares masculinos e femininos com AUC de 0,97) [22].

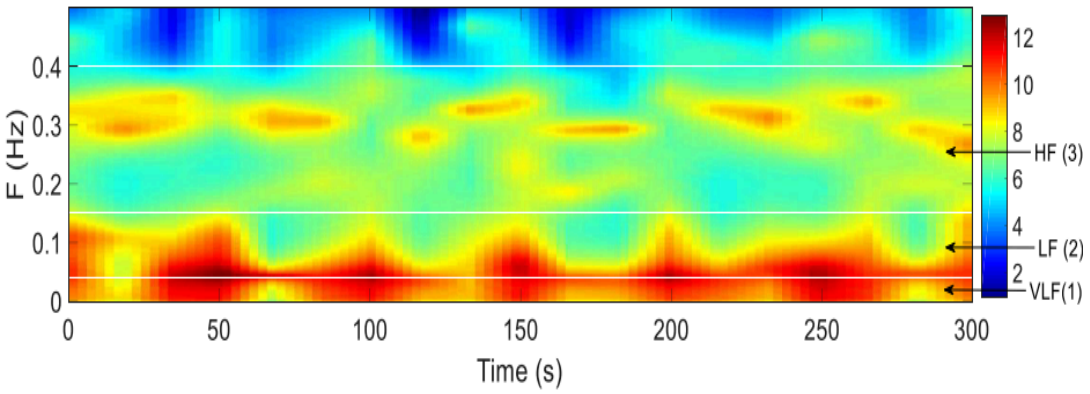


Figura 4 - Exemplo de imagem Power Spectral Density. Neste caso, a variável analisada foi a variabilidade da frequência cardíaca [23].

O desafio atual é permitir que a IA possa prever uma sépsis com 12 a 24 horas de pré-aviso; noutras patologias (como a retinopatia da prematuridade e a displasia broncopulmonar) esse intervalo pode ser mais longo.

3.2.3 Monitorização Não-Invasiva

A pele do RN prematuro é mais frágil do que a do RN de termo, constituindo assim uma população de interesse na monitorização de sinais vitais sem contato. O método tradicional utiliza transdutores e elétrodos com adesivos, os quais, apesar de adaptados até um certo limite às características desta pele, são frequentemente causa de lesão com abrasão e perda de continuidade cutânea, abrindo uma via para a infeção [24], [25].

Várias abordagens podem ser utilizadas para evitar este contato [24], [25], [26], [27]:

- indução magnética - deteta movimentos cardíacos e respiratórios, mas é altamente influenciado pelo próprio movimento do RN;

- termografia - caracterização das diferenças térmicas induzidas pela pulsação nos vasos mais superficiais; são muito dependentes de imobilização, têm de ser adaptadas e calibradas para cada RN (Figura 5);

- radar – deteta movimentos da caixa torácica e assim pode deduzir, através do efeito de doppler, sobre a FC e FR, mas é influenciado pelos movimentos; também não está esclarecido se é um método seguro pelos prováveis efeitos biológicos;

- lidar – variação do método anterior, mas com recurso a luz; tem algumas das limitações anteriores, mas será mais seguro nos efeitos secundários biológicos.

Uma opção mais recente visa utilizar as imagens captadas por câmaras digitais que trabalham com sensores que detetam o espectro visível e infravermelho (400-1000nm). Utilizou uma amostra de 30 RNMBP e 426 horas de gravação e posteriormente comparou imagens e dados convencionais de elétrodos, processados

numa CNN, com bons resultados, não influenciados pela etnicidade / cor de pele [26].

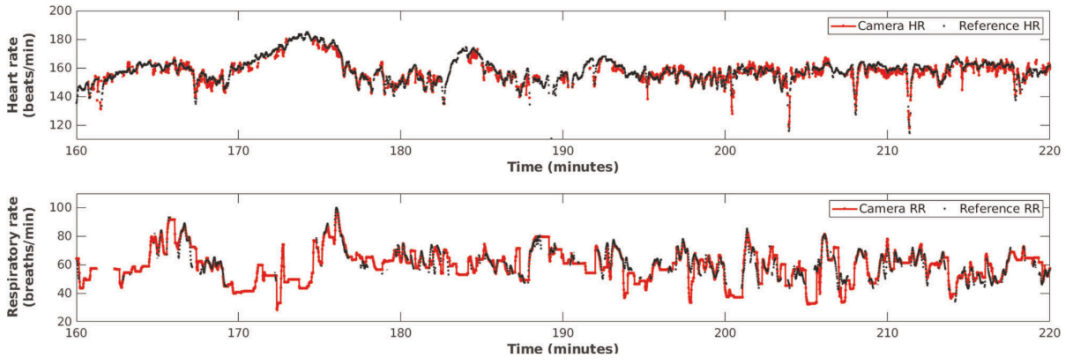


Figura 5 – Comparação entre a FC (elétrodos) e a FC calculada [26]

Como principal desvantagem de todas estas opções deve ser referida que dependem da colocação de uma câmara por cima ou diretamente em contato com uma incubadora. Ora as paredes da incubadora, apesar de transparentes, podem estar embaciadas pela condensação. As incubadoras mantêm uma atmosfera com humidade > 45%, sendo que em idades gestacionais mais baixas (<27 semanas), devem manter valores 70-80%. Ora esta atmosfera também está aquecida, podendo ter temperaturas de ar em torno dos 36-37ºC. Também a utilização de fototerapia com luz azul no tratamento da hiperbilirrubinémia pode influenciar a captação de imagens como se exemplifica na Figura 6.

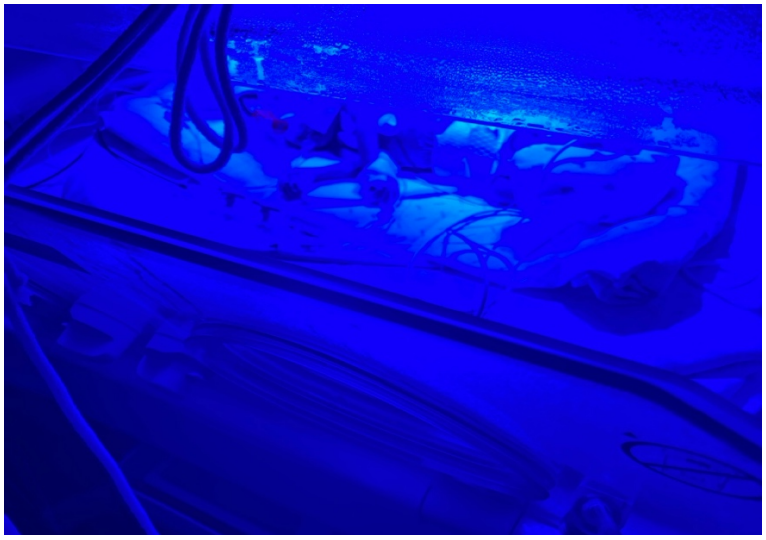


Figura 6 – RN prematuro em fototerapia (luz azul – 460nm)

E, por último, de acordo com as orientações dos cuidados centrados no desenvolvimento, o RN deve estar "aconchegado" com material de conforto, dificultando a obtenção de imagens, como se observa na Figura 7.

Este posicionamento do RN com material de conforto visa reproduzir o ambiente e posição naturais, ou seja, aquelas que são experienciadas durante a gestação.

Figura 7 – RN PT, 24 semanas, posicionado com material de conforto [28]

3.2.4 IA na Imagiologia Neonatal

Um RN necessitará, ao longo da sua estadia na UCIN, de múltiplas aquisições de imagem, seja com radiação ionizante, seja com ecografia ou ressonância magnética.

No caso da radiação ionizante, cuja utilização está a ser cada vez mais limitada pelos efeitos secundários, o papel da interpretação por IA será menos importante.

A ecografia permite a construção de uma imagem através do processamento dos diversos efeitos da passagem de ondas sonoras imperceptíveis ao ouvido humano. Esta tecnologia tornou-se imprescindível em cuidados intensivos neonatais: rápida, relativamente simples de manipular, pode ser feita à cabeceira do doente, apenas implica uma curva de aprendizagem razoável. No momento atual, podemos referir a ecografia cerebral, a cardíaca, a pulmonar, a abdominal e múltiplos procedimentos ecoguiados, como a colocação e visualização de cateteres centrais e a punção lombar ecoguiada.

Na ecografia cerebral (ecografia transfontanelar) já permite a identificação de estruturas ventriculares e calcular o índice ventricular e estimar o volume ventricular automaticamente a partir de imagens 2D [29], [30].

Na ecografia pulmonar permite não apenas a identificação de patologia fundamental como o pneumotórax, mas também calcular scores na doença de membrana hialina (deficiência em surfactante nos prematuros) praticamente em tempo real [31], [32]. Eventualmente, a mesma abordagem pode ser aplicada para elaborar scores

preditivos na displasia broncopulmonar (doença pulmonar crónica da prematuridade)[33]. Aliás, de acordo com algumas referências, todo o exame poderá ter a identificação de padrões característicos de patologias de forma completamente automatizada com recurso a IA [34], [35].

Na ecografia cardíaca permite identificar com precisão a presença de canal arterial. Uma das referências utilizou 66 RN dos quais foram extraídos 1145 clips de vídeo (661 com canal arterial presente, 484 sem), os quais foram processados numa CNN e atingiu uma sensibilidade de 0.80 (0.83–0.90), especificidade de 0.77 (0.62–0.92) e AUC de 0.86 (0.83–0.90)[36]. Eventualmente, poderá ser possível aplicar em neonatologia o modelo que permite a identificação automatizada de planos / vistas e estruturas validado para idade pediátrica [37].

Na radiologia convencional, permite a identificação de cateteres variados com razoável precisão: um estudo refere ter desenvolvido um algoritmo de CNN com base em 777 radiografias toraco-abdominais de RN, com uma precisão global de 0,97, quer se trate de um ou vários cateteres em simultâneo [38] (ver Figura 8).

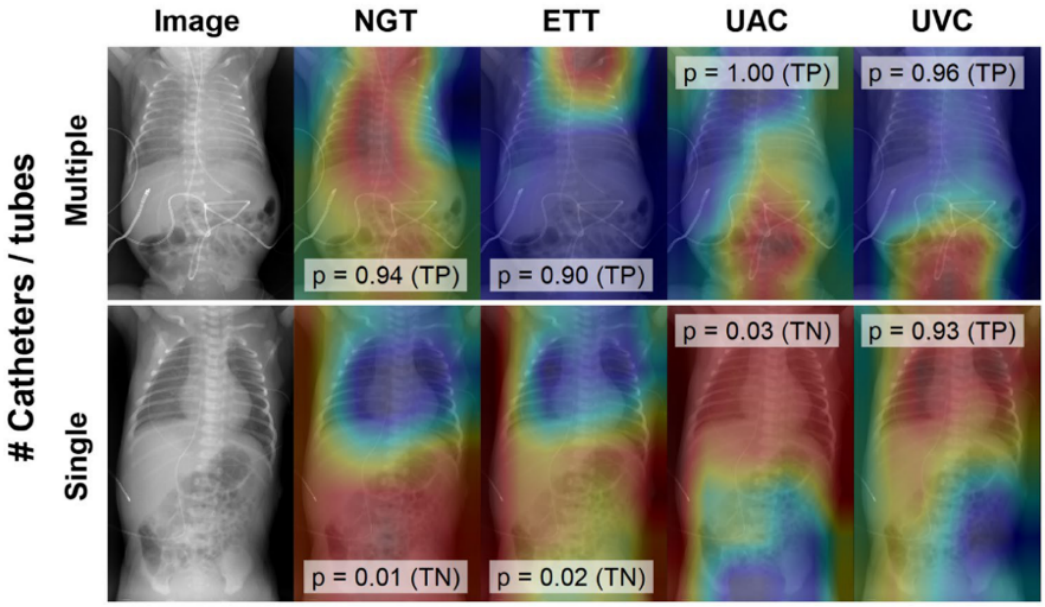


Figura 8 – Identificação de sonda nasogástrica (NGT), tubo endotraqueal (TET), cateter venoso umbilical (UVC) e cateter arterial umbilical (UAC) [38]

Na neuroimagem por ressonância magnética, aquela que mais anos tem de experiência com IA, a questão vai um pouco mais longe: num estudo foi possível a previsão de perfis de neurodesenvolvimento aos 2 anos com base em imagens obtidas entre as semanas 38 e 43 de idade corrigida e, assim, redireccionar cuidados e intervenções na janela de neuroplasticidade dos primeiros 24 meses de vida [39]. Esta perspetiva permitirá modificar e melhorar o perfil de neurodesenvolvimento.

3.2.5 Ventilação do RN Prematuro e IA

O RNMBP necessita, frequentemente, de ser ventilado de forma invasiva ou não invasiva. A primeira é a mais complexa e aquela que mais riscos contempla, sendo também implicada em diversas patologias e morbi-mortalidade.

A ventilação de um pulmão imaturo, disfuncional, sem surfactante e em crescimento é um risco assumido. Apesar de todas as abordagens de *lung protection ventilation, gentle ventilation, volume-targeted ventilation* e afins, a ventilação mecânica é o principal fator de risco para displasia broncopulmonar e a duração e severidade da ventilação correlaciona-se com mortalidade e morbilidade.

A integração da IA pode facilitar a decisão de progredir para a extubação, com base na avaliação de vários parâmetros nas duas horas anteriores e conseguiu uma AUC de 0,871 e sensibilidade de 70,1% e 90% de especificidade [40], [41] .

A aplicação mais urgente é a capacidade de ajustar quer os parâmetros, quer os modos de ventilação em uso num determinado RN. Apesar de cada quadro clínico e suporte ventilatório serem generalizáveis, a sua aplicação a cada RN (também globalmente generalizável por idade gestacional e peso de nascimento) é uma questão de medicina individualizada.

Assim, caberia à IA o ajuste a curto prazo dos diversos parâmetros ventilatórios através do processamento dos múltiplos dados fornecidos pelo ventilador. Esta abordagem foi já tentada, conseguindo prever o comportamento de alguns parâmetros ventilatórios com 1,5s de antecedência, como se observa na Figura 9 [41].

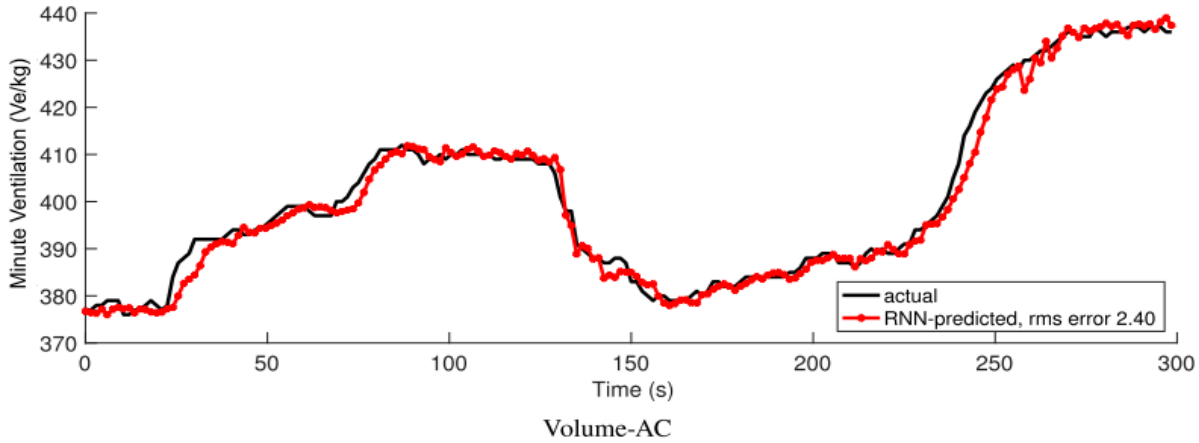


Figura 9 – Predição do volume minuto em ventilação AC-VG com base nos 5 ciclos anteriores. Volume minuto é o volume total de gás que entra (ou sai) do pulmão por minuto. É igual ao Volume Corrente (VC) multiplicado pela frequência respiratória (f): MV = VC x f [41].

3.2.6 Retinopatia da Prematuridade

A Retinopatia da Prematuridade (ROP) é uma doença vasoproliferativa da retina que é uma das principais causas de deficiência visual e cegueira infantil em todo o mundo. Caracteriza-se por uma fase inicial de degenerescência microvascular da retina, seguida de neovascularização, que pode levar ao descolamento da retina e à perda visual permanente [42]. Afeta entre 30 e 50% dos RNMBP, dos quais 25 a 30% progridem para alterações visuais graves, com disfunção da retina e mesmo cegueira. [42].

A utilização de IA na fundoscopia já provou ser superior ao oftalmologista, ao conseguir distinguir entre um fundo ocular masculino e feminino [22].

A ROP implica um acompanhamento regular por oftalmologista dedicado, frequentemente apenas disponível em centros diferenciados. Existem já diversos projetos que recorrem a tele-oftalmologia para evitar deslocar os RN. Recentemente, uma metanálise referenciou 18 artigos que comparam a utilização de algoritmos IA para processar imagens recolhidas por RetCam©, não apenas para identificar RN em risco de ROP, mas também para detetar e classificar a ROP por estadios e referenciar os que necessitariam de tratamento [43].

Assim, a integração da IA no diagnóstico e caracterização da ROP é lógica. Um exemplo, entre vários, pode ser encontrado neste trabalho [44]. A proposta é permitir a utilização do sistema por vários locais, baseando-se na *cloud computing*, como se ilustra na Figura 10.

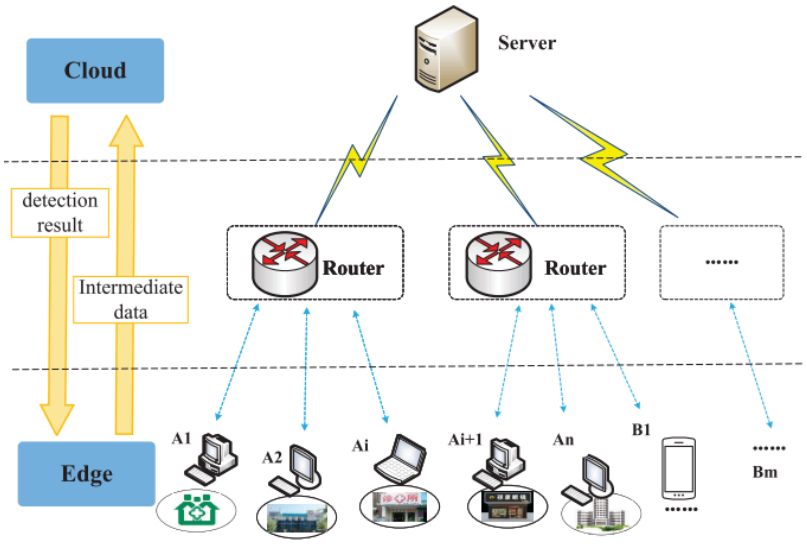


Figura 10 – Diagnóstico remoto de ROP com *deep learning* e *cloud computing*. Neste sistema, apenas os dados pré-processados das imagens da RetCam© são enviados para uma *cloud* [44].

A opção de incluir a IA no diagnóstico e estadiamento da ROP revelou melhores resultados, ao permitir uma classificação baseada em quantificadores, indo além da classificação clínica (baseada em vários critérios, mas ainda assim subjetiva)[45].

O desafio atual está na capacidade de prever a progressão para ROP, estabelecendo um prognóstico antes das 45 semanas de idade pós-menstrual. Este artigo utilizou 7033 imagens de retina correspondentes a 725 RNMBP para treino e 763 imagens de 90 RNMBP para validação, bem como 46 outras características inerentes aos RN. Estas foram processadas por duas redes neuronais, em que uma primeira detetava a ocorrência de ROP e uma segunda classificava a severidade. No final, o sistema total permitiu identificar RNMBP com elevado risco de desenvolver ROP [46].

Assim, a integração da IA na identificação de fatores de risco, deteção e seguimento da ROP é francamente superior à estratégia atual. Permite ainda menos deslocações a grandes centros para acompanhamento e transferências mais precoces para a área de residência, adicionando mais uma vantagem na utilização de sistemas baseados em IA.

3.2.7 Monitorização da Dor

A dor constitui o 5º sinal vital mas, infelizmente, não pode ser continuamente monitorizada como a FC ou FR. Numa UCIN, os RN não podem colaborar na quantificação da mesma como se pode fazer em idade pediátrica com uma régua de dor, por exemplo. Os RN são frequentemente expostos a dor aguda, repetitiva e crónica em cuidados intensivos, quer se trate de procedimentos, cuidados ou terapêuticas. Os RNMBP, especialmente abaixo das 30 semanas de gestação, estão expostos a 10 a 15 procedimentos dolorosos por dia, numa fase em que a dor não é inócua no neurodesenvolvimento. Um maior número de procedimentos dolorosos tem sido associado a um desenvolvimento neurológico patológico com maturação cerebral alterada e mesmo a um crescimento somático pós-natal abaixo do esperado [47].

O principal desafio consiste em identificar estímulos dolorosos ligeiros / borderline que podem passar despercebidos e não ter tratamento adequado. Este problema é tanto mais complexo quanto menor a idade gestacional, pois os RNMBP mais imaturos podem nem conseguir apresentar choro sempre que estão incomodados. A resposta fisiológica habitual com alteração no padrão de FC, FR e restantes parâmetros fisiológicos é muitas vezes acompanhada de expressões faciais e movimentos dos membros. Toda esta informação deve ser percecionada pela equipa médica/enfermagem, num ambiente já de si lotado em estímulos e quebras de atenção.

Um dos trabalhos mais recentes nesta área propõe uma solução mais simples, ao utilizar apenas imagens em tempo real. Estas são processadas por uma CNN com o modelo YOLO© (You Only Look Once), sendo que os alertas são depois

manualmente verificados. O modelo foi treinado com recurso a grandes *datasets online* (VGGFace, COPE, MNPAD) e depois testado em ambiente local, com uma deteção 25% superior e uma precisão de 77% [47], [48].

Uma abordagem ligeiramente diferente utilizou o recurso FaceReader©, com resultados muito promissores, distinguindo entre estímulos não dolorosos, moderamente dolorosos e extremamente dolorosos, com uma correlação de 0,84 ($p<0,001$) quando comparado com a avaliação humana [49]. O esquema global de processamento está ilustrado na Figura 11.

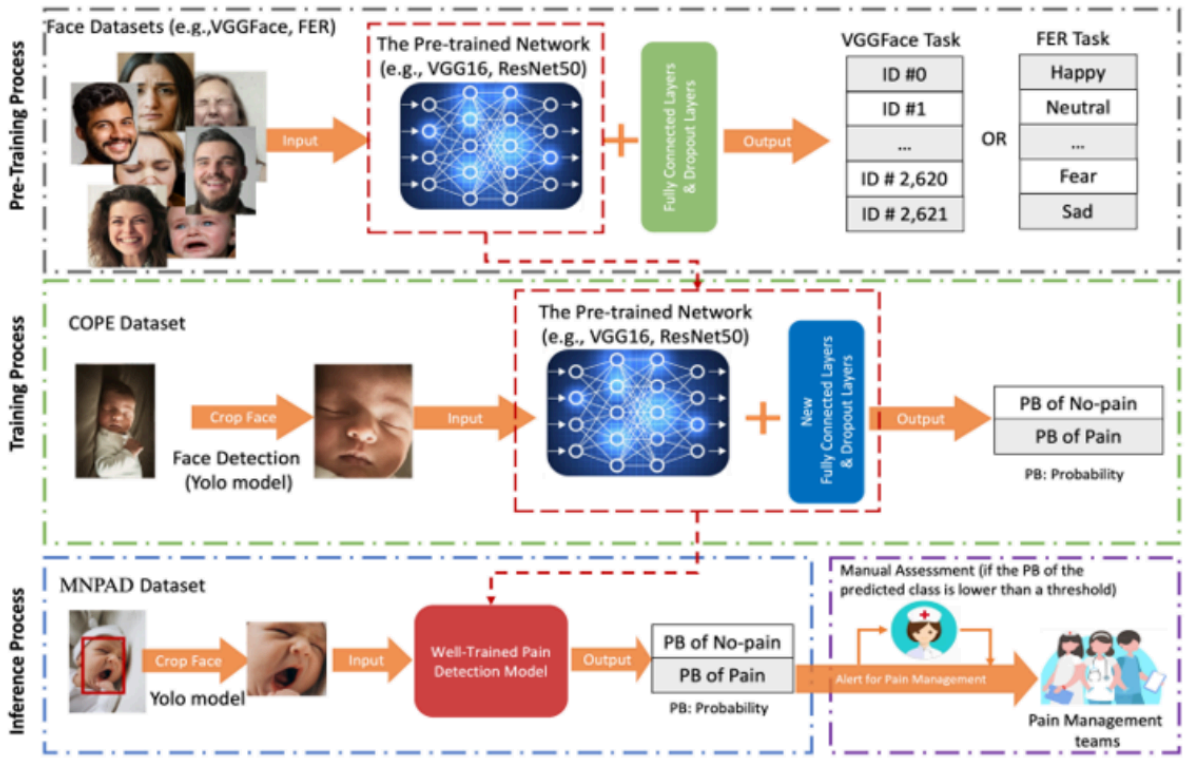


Figura 11 – Modelo proposto [47]

Uma abordagem distinta passa pela análise do choro. O processo inicia-se com a recolha de amostras de clips de som, divididos em dor / não dor, que são transformados em espectrogramas. Este método conseguiu uma precisão de 91,2% com AUC 0,91 [50]. Esta abordagem seria mais simples do ponto de vista da implementação, mas implica que haja choro, o que não ocorre em RN muito imaturos / extrema prematuridade e em RN ventilados, o que é uma limitação séria.

3.3 A IA na Eficiência e Qualidade

A Qualidade e a Eficiência são cada vez mais aspetos incontornáveis na medicina. A Neonatologia beneficia com a implementação de IA ao permitir uma medicina individualizada, preditiva, baseada em evidência, mas também ajustada às características do local e com uma utilização racional dos meios disponíveis.

Um trabalho recente explorou a capacidade preditiva de vários classificadores fundidos num único algoritmo - Logistic Regression, ExtraTrees, Random Forest, KNN, Support Vector Classifier, AdaBoost, GradientBoosting, XGBoost, CatBoost e Random Forest – para estimar a duração do internamento numa UCIN. Utilizou as notas clínicas da alta de 453 RN, das quais foram extraídas 12 features (peso, idade gestacional, idade da mãe, tipo de parto, nível de cuidados exigidos, entre outras variáveis). No final, classificava em duração de internamento curta (até 25 dias) e longa (superior a 25 dias). Globalmente, atingiu uma precisão de 0,96 e um AUC de 0,95 [51]. O esquema global de processamento está ilustrado na Figura 12.

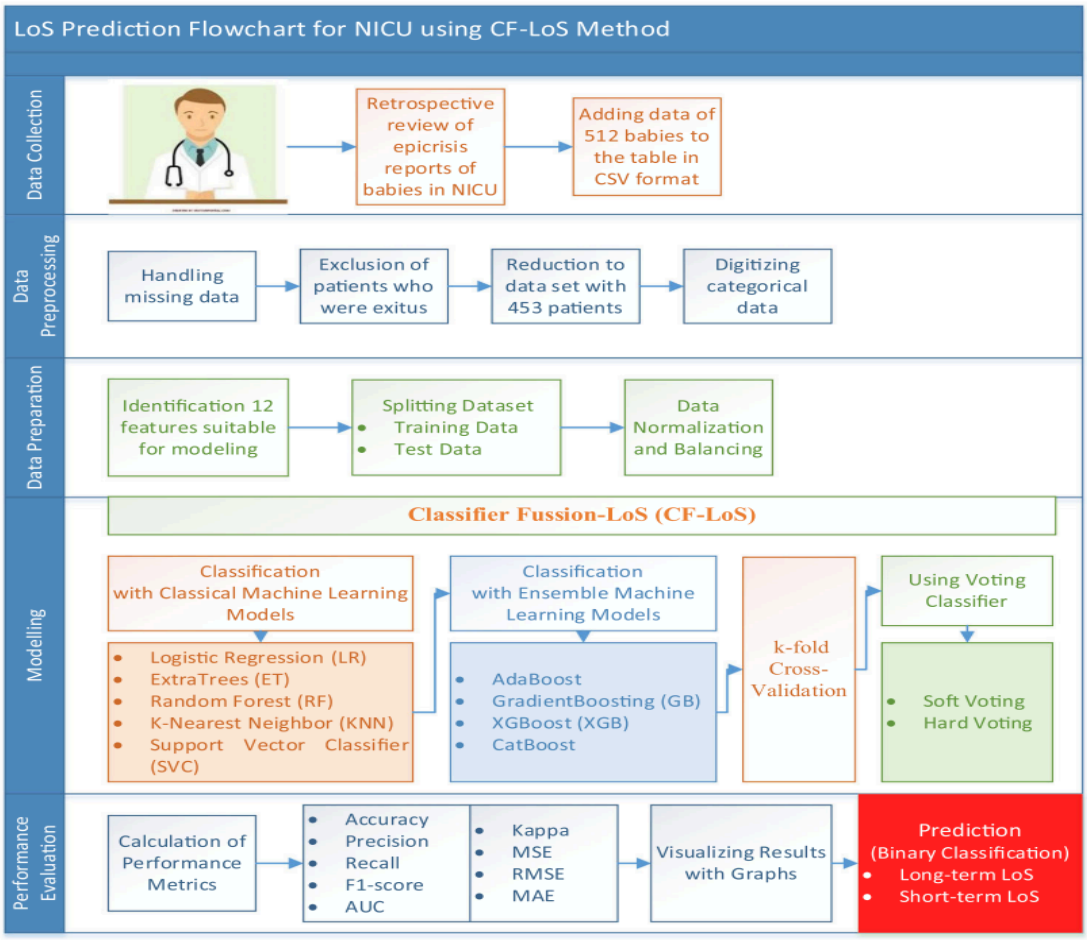


Figura 12 – Flowchart Classifier Fusion para predição de Length of Stay [51].

Este tipo de abordagem é fundamental na perspetiva da alocação de recursos, quer físicos, quer humanos, numa unidade hospitalar. Apesar das lições aprendidas com a pandemia SarsCov2, ainda não temos uma gestão nacional (ou regional) dos recursos das UCINs. Ao estimar a duração de internamento, permite também ajustar futuras necessidades desses mesmos recursos, dimensionamento dos serviços e fluxos de transferências inter-hospitalares.

Seria interessante a aplicação de um modelo semelhante ao anterior numa perspetiva global, nacional (ou regional).

Ao longo do ano atual (2025) temos assistido a vários constrangimentos na acessibilidade aos cuidados de saúde urgentes, especialmente na vertente Obstétrica. Deve ser equacionada a utilização de modelos de IA para a gestão das grávidas e, eventualmente, também para os RN com necessidade de cuidados intensivos (UCIN), sendo mais eficazmente referenciados.

4 APLICAÇÃO DE IA NA UCIN DA ULS

4.1 O Problema da Nutrição Adequada no RN

Pretende-se aplicar a IA de uma maneira pragmática na avaliação e ajuste nutricionais de RN internados numa UCIN.

A nutrição do RN é fundamental para um adequado crescimento, para além de ter de manter todo o metabolismo basal e ainda fazer face a todas as situações de stress como o trabalho respiratório aumentado, o combate a infeções e o trabalho cardíaco aumentado na transição da circulação fetal para a circulação autónoma.

Numa fase inicial, o RN prematuro não tem capacidade de digerir a totalidade do aporte alimentar necessário a todas estas funções. Necessita, por isso, de um aumento progressivo de acordo com a sua tolerância.

Durante o período de transição, necessitamos de diversas estratégias para resolver este problema:

- No primeiro dia a dependência de Nutrição Parentérica Total (NPT) endovenosa é total; depois a mesma vai sendo retirada progressivamente, substituída pela nutrição entérica. Apesar de parecer um processo linear, frequentemente progride por avanços e recuos intermitentes. A sépsis e a enterocolite necrotizante podem levar a um retrocesso total; e a tolerância gástrica e intestinal não são lineares[52], [53].
- O início da alimentação enteral com Leite Materno (LM) é uma estratégia comprovada para diminuir o risco de ambas as complicações referidas acima. De acordo com a evidência científica, o intestino do prematuro deve receber LM preferencialmente, desde o primeiro dia, mesmo que em quantidades mínimas (alimentação enteral mínima / "*gut priming*") [54], [55]. Esta prática permite o estabelecimento de colónias bacterianas favoráveis (derivadas do leite materno), diminuindo assim o risco de complicações major (enterocolite necrotizante, por exemplo) e a incidência de doença alérgica futura.
- A velocidade de progressão também tem sido alvo de alguma controvérsia.
- Adicionalmente, esta situação possui maior complexidade: desde o primeiro dia de vida (D0), o volume total de líquidos (aporte hídrico total) é crescente, variando ainda com o ganho ou perda ponderal, a idade gestacional, o peso de nascimento e alguns fatores de risco pré-natais [54], [55].

A evolução global está ilustrada na Figura 13, tendo como exemplo um RN de 27 semanas.

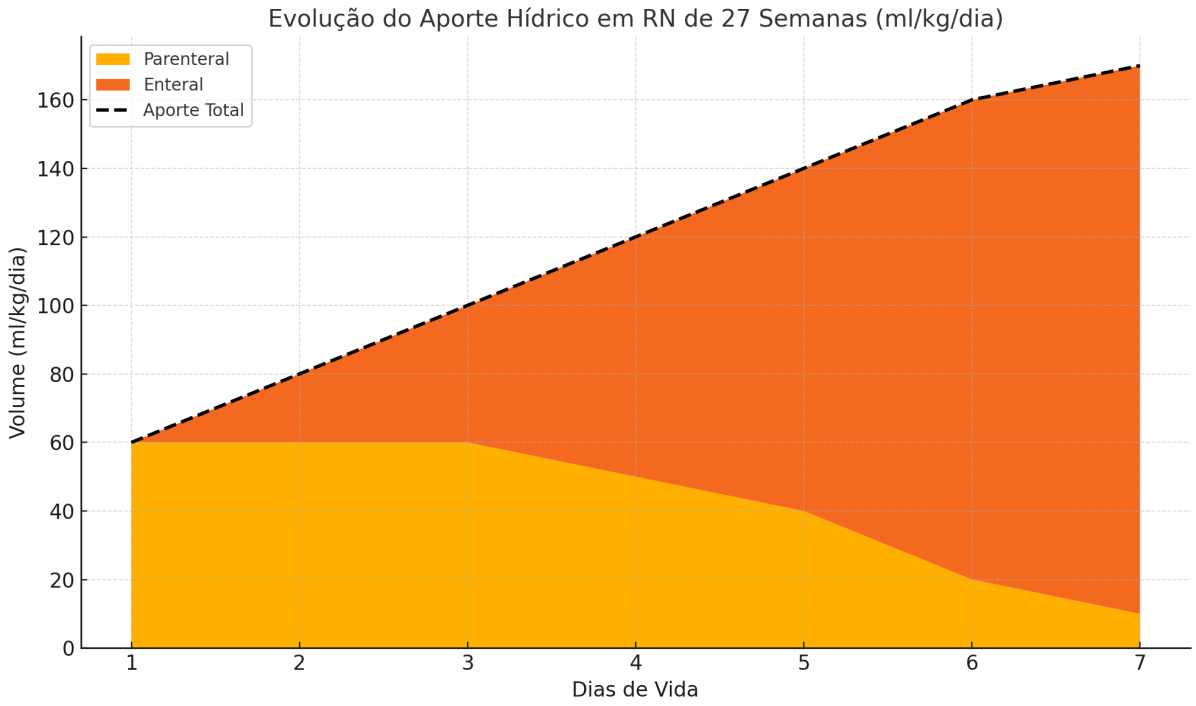


Figura 13 – Evolução do Aporte Nutricional em Volume ao longo da primeira semana de vida

Se a questão do volume total e do tipo de nutrição é complexa, não devemos esquecer os aportes calóricos e proteicos, pois serão estes que vão permitir um metabolismo e crescimento adequados.

A questão fundamental vai para além do ajuste limitado aos constituintes: volume e densidade calórica, proteínas, lípidos e hidratos de carbono, bem como a relação entre si [52], [53], [54], [55]. Numa UCIN, a atenção da equipa clínica pode frequentemente ser dispersa pelos diversos RN internados, motivada pela gravidade das patologias e também pela taxa de ocupação. Ora, se a atenção está dividida e tende a focar-se no mais importante, os ajustes nutricionais mais finos ficam para uma segunda oportunidade, por vezes inexistente. Ao fim de alguns dias, o diferencial torna-se importante.

Então como solucionar este problema?

Podemos equacionar a utilização de IA para, de forma automatizada, avaliar a adequação dos aportes nutricionais de cada RN.

De uma maneira geral, pretende-se extrair a informação das notas do diário clínico, comparar com os diversos protocolos e standards existentes e, no final, elaborar uma proposta.

Seguramente que teremos de abordar as diversas fases deste processo.

4.2 Extração de Elementos Clínicos

Então para que se possa melhorar todo o processo, necessitamos de obter os dados reais do RN em questão. Estes podem ser dados biométricos, demográficos, somatométricos, nutricionais, etc. E estão contidos em texto livre no diário clínico.

As opções atuais em uso na UCIN implicam a recolha manual dos dados e a sua introdução num Excel© para calcular os aportes nutricionais reais que o RN tem naquele momento. Se bem que este já permitiu ganhos em exatidão ao fazer cálculos automaticamente, é ainda um processo manual. Este método é moroso, sujeito a erros de transcrição e distração por divisão da atenção.

E assim chegamos ao ponto fundamental de todo o processo: a extração de dados clínicos.

Esta extração deve ser rápida e exata. Mas este é todo um problema por si mesmo, pois é a etapa mais complexa e falível, muito condicionada pelos métodos utilizados e pelas fontes desses dados – o texto clínico.

É também um dos campos em rápida expansão na IA na Medicina.

4.2.1 Natureza do Texto Clínico e Desafios da Extração

Os textos médicos podem ser classificados em três níveis de organização: estruturados (ex.: resultados laboratoriais), semi-estruturados (ex.: notas de evolução ou diários clínicos com notas organizadas por sistemas) e não estruturados (ex.: consultas em texto livre). Os dados estruturados são facilmente pesquisáveis, mas muitas vezes não captam as subtilezas clínicas do contexto ou tendem a omitir informação clínica implícita ou contextual. O texto não estruturado varia no sentido inverso: contexto clínico com muita informação, mas de difícil processamento para extração sistemática [56].

A extração de conceitos clínicos deve ainda considerar a variabilidade na terminologia, abreviaturas, negações, referências temporais e contexto. As notas médicas estão repletas de acrónimos para os quais frequentemente não existe padronização, mesmo dentro da mesma instituição.

4.2.2 Abordagens Baseadas em Regras e Padrões

Os primeiros trabalhos dependeram fortemente de sistemas baseados em regras, incluindo *Regular Expressions* (RegEx) e scripts específicos de domínio — conjuntos de regras desenvolvidos manualmente por especialistas clínicos, adaptados ao vocabulário e à estrutura documental de um determinado serviço ou instituição. Ferramentas como o cTAKES e o MetaMap fazem o mapeamento de texto para ontologias médicas padronizadas como a UMLS ou a SNOMED CT. Estes sistemas são concebidos para identificar termos médicos em texto livre — como diagnósticos, sintomas ou procedimentos — e associá-los a conceitos estruturados em ontologias [57], [58], [59], [60].

Embora eficazes na identificação de conceitos bem definidos, estas ferramentas não estão aptas a processar expressões ambíguas ou específicas de uma instituição, que não estão incluídas em vocabulários pré-definidos. Além disso, os conjuntos de dados neonatais são escassos.

A correspondência de padrões continua útil em cenários específicos, nomeadamente quando se trabalha com documentação em formato padrão, como folhas de registo de cuidados intensivos ou diários clínicos neonatais. Contudo, as limitações são evidentes quando a linguagem é inconsistente ou quando é necessária alguma descodificação semântica e contextual.

4.2.3 Sistemas Baseados em Estatística e Aprendizagem Automática

A introdução destas abordagens representou um avanço. Conditional Random Fields (CRFs) e Support Vector Machines (SVMs) tornaram-se amplamente usados para Named Entity Recognition (NER) em texto clínico. Estes modelos identificam e classificam o texto em categorias clínicas pré-definidas, como diagnósticos, procedimentos e medicação [57], [61].

Em cuidados neonatais, modelos baseados em CRF foram aplicados para identificar "escalada de suporte ventilatório", "desmame de oxigénio" ou "enterocolite necrosante" em notas clínicas. Se existir treino para aprender dependências contextuais, os CRFs podem distinguir se a referência a "ventilação" diz respeito a ventilação mecânica invasiva ou ventilação não invasiva.

Do mesmo modo, SVMs treinadas em conjuntos de dados anotados podem diferenciar entre intervenções terapêuticas e observações diagnósticas. Em neonatologia, foram usadas experimentalmente para classificar expressões como "escalar suporte de ventilação não invasiva" ou "suplementação de oxigénio" como intervenções ativas. Também foram testadas para distinguir diagnósticos atuais de outros passados, recorrendo a pistas no texto.

Contudo, estes modelos apresentam dificuldades em incorporar contexto para além do nível da frase. São eficazes quando treinados em contextos específicos, mas a sua performance decai quando aplicados a novos conjuntos de dados com variação linguística significativa. Além disso, requerem características específicas e grandes quantidades de dados anotados manualmente — em neonatologia não tem sido fácil encontrá-los.

4.2.4 Aprendizagem Profunda e Modelos de Linguagem Contextuais

As técnicas de *deep learning*, especialmente aquelas baseadas em *Recurrent Neural Networks* (RNNs) e redes *Long Short-Term Memory* (LSTMs), representaram um avanço importante. Estes modelos permitiram uma melhor gestão das dependências temporais e das relações não lineares no texto clínico. No entanto, acabaram por ser ultrapassados por arquiteturas baseadas em *transformers*.

Em 2018, a Google© introduziu o *Bidirectional Encoder Representations from Transformers* (BERT), que revolucionou o processamento de linguagem natural. As suas variantes clínicas, como o *ClinicalBERT* e o *BioBERT*, continuam a demonstrar bons resultados em tarefas como o reconhecimento de entidades nomeadas (Named Entity Recognition - NER) e a extração de relações semânticas. Estes modelos utilizam mecanismos de atenção para considerar o contexto a partir de ambas as direções do texto, oferecendo uma interpretação mais precisa — algo essencial na prática clínica.

Em ambiente neonatal, estes modelos conseguem extrair informação detalhada como o tipo e o momento de início do suporte respiratório ou o estado nutricional. Por exemplo, o *ClinicalBERT* tem sido utilizado para identificar fatores de risco em recém-nascidos prematuros, reconhecendo pistas subtis em notas clínicas associadas a desfechos como a displasia broncopulmonar [57], [62], [63], [64].

Ainda assim, mesmo modelos treinados especificamente para determinado ambiente clínico, como o RoBERTa no âmbito Cirurgia Cardiotorácica, tem uma performance inferior ao de vários LLMs, apesar de ser uma solução com necessidade de saída de dados do ambiente hospitalar [56].

A capacidade de compreender o contexto permite que o modelo distinga entre história clínica prévia e condições atuais — uma diferenciação crítica em notas clínicas neonatais diárias. Ainda assim, estes modelos necessitam frequentemente de *fine-tuning* para se adaptarem aos padrões linguísticos específicos das UCINs.

4.2.5 Integração de MedSpaCy e Pipelines Híbridos

No contexto dos modelos baseados em linguagem, ferramentas como o *MedSpaCy* funcionam como complemento aos modelos *transformer*, adicionando camadas de interpretação semântica específicas para o domínio clínico. Estas incluem: regras explícitas para reconhecimento de padrões clínicos, identificação da secção do documento onde cada expressão aparece e deteção de negações — tudo adaptado à estrutura e terminologia da documentação médica.

Além disso, o algoritmo *ConText*, incluído no *MedSpaCy,* permite identificar se determinada condição está negada ou se está apenas a ser vigiada. Esta combinação entre modelos estatísticos e regras explícitas torna o pipeline mais robusto, modular e explicável.

O *MedSpaCy* é especialmente eficaz quando combinado com reconhecedores de entidades como o *spaCy* ou o próprio *ClinicalBERT*, formando pipelines híbridos que integravam o melhor dos dois mundos até há pouco tempo [62], [64], [65].

4.2.6 Large Language Models na Prática Clínica – e porque não?

A introdução de Large Language Models (LLMs) como o GPT©, LLaMA© e PaLM© representa uma mudança de paradigma. Ao contrário dos modelos anteriores, os LLMs podem ser adaptados a tarefas diversas através de simples *prompting*, sem necessidade de treino adicional específico para cada tarefa.

Cada família de modelos oferece vantagens distintas: o *PaLM* (*Pathways Language Model*, da Google) disponibiliza capacidades multimodais e multitarefa integradas em ferramentas como o Bard; o *GPT* (*Generative Pre-trained Transformer*, da OpenAI) alimenta aplicações como o ChatGPT e o Codex; e o *LLaMA* (*Large Language Model Meta AI*, da Meta) disponibiliza modelos eficientes e *open-source*.

Estudos recentes demonstraram que o GPT-3 e modelos similares são eficazes em tarefas como sumarização de registos clínicos, geração de instruções de alta e classificação *zero-shot* de eventos médicos. Contudo, existe o risco de alucinação e reprodução de enviesamentos dos dados de treino.

Recentemente, uma publicação comparou a performance de múltiplos LLMs na extração de dados clínicos, com classificação binária e extração de entidades (NER), em que a maioria teve uma pontuação excelente, como se ilustra na Figura 14. Podemos mesmo dizer que estes são, pelo menos, não inferiores ao operador humano [57], [66], [67], [68], [69].

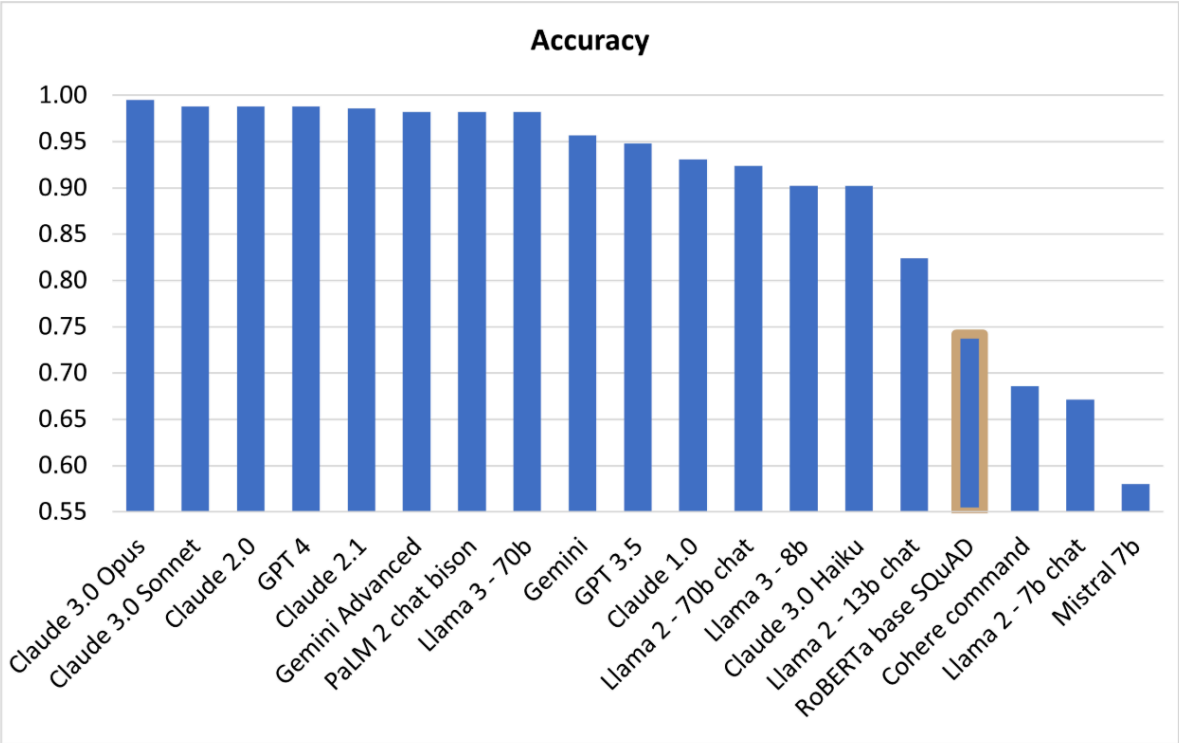


Figura 14 – Comparação na performance de extração de dados clínicos de múltiplos modelos baseados em LLM com a referência BERT (RoBERTa SQuAD, já previamente treinado para o efeito) [56]

A aplicabilidade dos LLM na medicina é cada vez mais ampla: na decisão clínica e otimização de algoritmos de decisão; na codificação clínica; na geração de resumos clínicos; na tradução de documentos atendendo ao domínio específico; na educação e formação médicas; na saúde mental como suporte a pacientes e ainda na criação de inquéritos a pacientes [70]. Todos estes aspetos estão exemplificados na Figura 15.

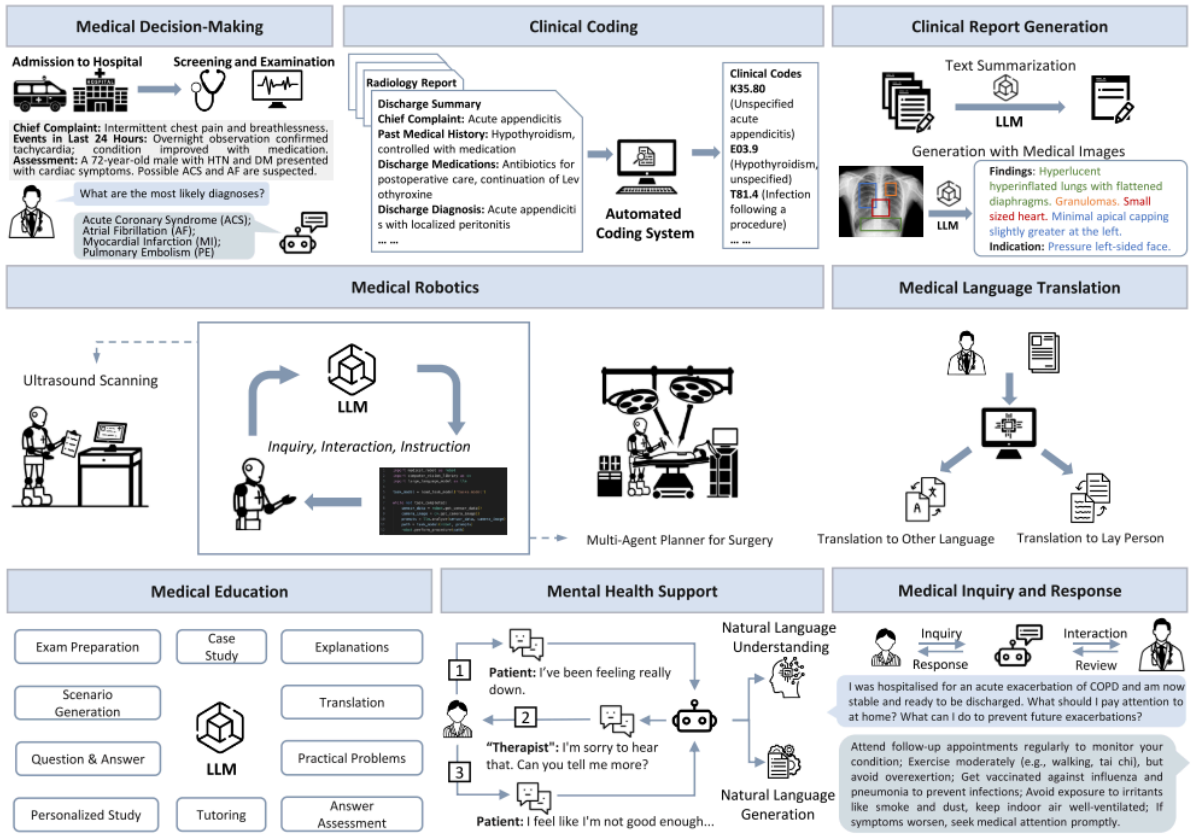


Figura 15 – LLMs na medicina em geral [70].

5 LARGE LANGUAGE MODELS

5.1 Introdução

O Processamento de Linguagem Natural (*Natural Language Processing* - NLP**)** pode ser definido como o campo da inteligência artificial que estuda como os sistemas computacionais podem compreender, processar e gerar linguagem humana. Engloba técnicas clássicas baseadas em regras, métodos estatísticos e, mais recentemente, modelos de aprendizagem profunda [71].

Dentro deste campo emergem os Modelos de Linguagem de Grande Escala (*Large Language Models* - LLMs**)**, uma subcategoria de modelos de NLP caracterizados por três elementos centrais:

- Escala — treinados com grandes quantidades de texto, possuem milhares de milhões de parâmetros;
- Arquitetura Transformador *(Transformer)* — permitem captar dependências de longo alcance através de mecanismos de Atenção;
- Capacidade generalista — funcionam como modelos de propósito geral, adaptáveis a múltiplas tarefas (tradução, sumarização, resposta a perguntas, geração de código, conversa); como principal característica apresentam a capacidade de gerar texto e também interpretar.

Enquanto sistemas de NLP tradicionais eram treinados para tarefas específicas (ex.: um classificador de sentimentos ou ainda NER), os LLMs são pré-treinados em conjuntos muito abrangentes e completos de dados, podem ser ajustados *(fine-tuned)* ou simplesmente instruídos via *prompting* para desempenhar diferentes funções [71], [72].

5.2 Arquitetura Básica dos LLMs

5.2.1 Transformer

Um *Transformer* é uma rede neuronal que identifica o contexto de dados sequenciais e gera novos dados a partir destes. Estão concebidos para compreender o contexto e o significado através da análise da relação entre diferentes elementos, e baseiam-se no mecanismo de atenção para o fazer.

Antes dos modelos *Transformers*, outros modelos sequenciais como RNNs e LSTMs processavam *tokens* um a um, dificultando a captura de dependências longas (ex.: início e fim de uma frase longa). Os Transformers introduziram o mecanismo de atenção, permitindo que cada palavra *(token)* pudesse "olhar" para todas as outras simultaneamente, independentemente da distância no texto. Esse design elimina a recorrência e melhora o desempenho em tarefas complexas de NLP. Esta mudança paradigmática está na base dos LLMs modernos (GPT, BERT, LLaMA, PaLM, etc) [73].

O Transformer processa texto em etapas sucessivas, "transformando" frases em representações vetoriais manipuláveis matematicamente - ver Figura 16 e Figura 17.

Figura 16 – Esquema Transformer [73]

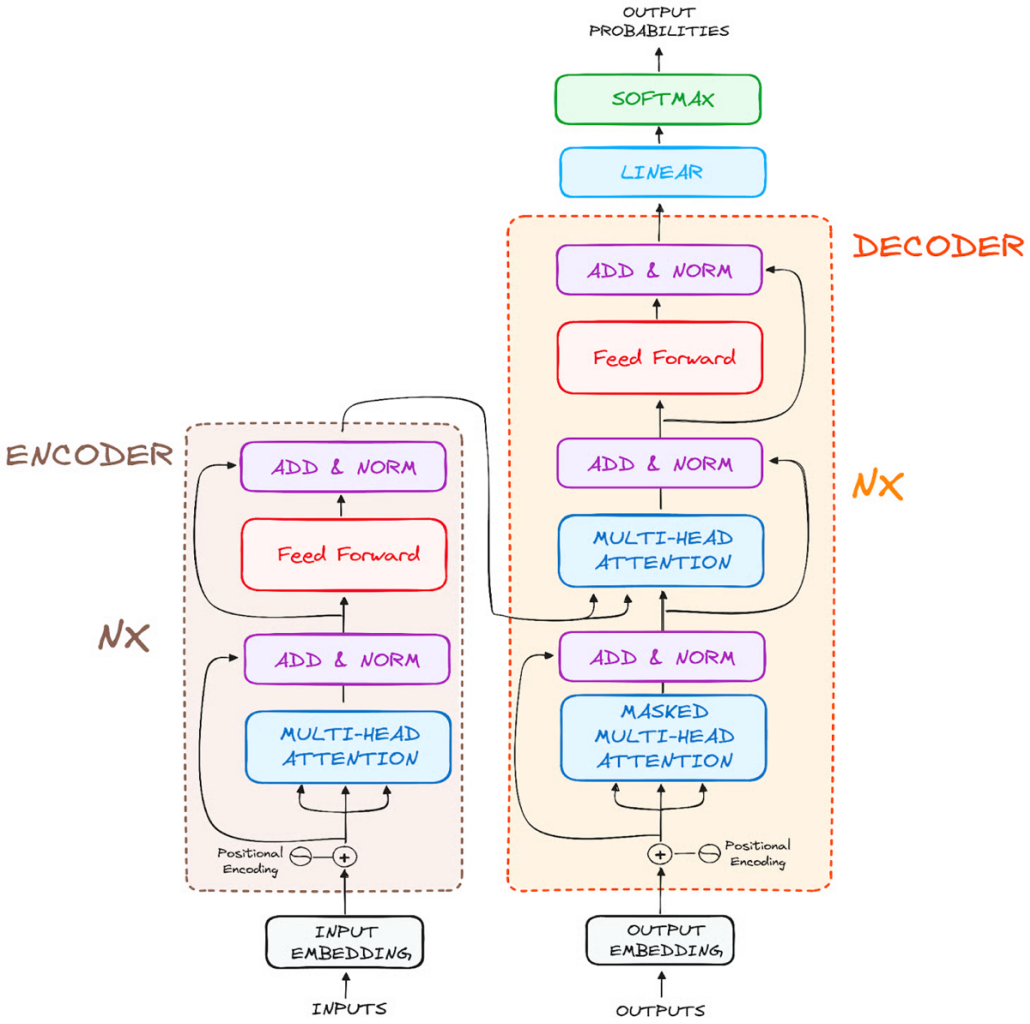


Figura 17 – Arquitetura global Encoder-Decoder [73]

a) Tokenização e Embedding

O texto é convertido em *tokens*, ou seja, em unidades fundamentais de pequena dimensão que podem ser processados individualmente:

Exemplo → "My Yellow Cat" → `[My] [Yellow] [Cat]`

Cada *token* é posteriormente representado (*Embedding*) por um vetor num espaço virtual de dimensão fixa. Através do embedding conseguimos representar relações e significados de dados originalmente não-numéricos (aqui considerámos "apenas" 6 dimensões):

My → `[0.2, 0.7, -0.1, 0.5, 0.3, 0.0]`

Yellow → `[0.9, -0.2, 0.4, 0.1, -0.5, 0.6]`

Cat → `[0.3, 0.8, 0.2, -0.4, 0.9, 0.1]`

Ou seja, cada palavra / elemento (token) foi convertida num vetor, os quais são aprendidos durante o treino e captam também semântica: palavras semelhantes ficam próximas no espaço (virtual) vetorial [74]. Cada modelo terá embeddings

próprios, e mesmo versões sucessivas do mesmo terão embeddings diferentes, apesar de poderem ter sido treinados com tokenizadores iguais. Em resumo, "cat" terá um embedding diferente no GPT5 e no Gemma, e também diferente entre os GPT 2, 3, 4 e 5...

Apesar de tudo, a lógica mantém-se – elementos (*tokens*) semanticamente próximos terão vetores "próximos" no espaço virtual hiperdimensional: "*dog*" e "*cat*" estarão mais "próximos" do que "*cat*" e "*house*", mas mais afastados do que "cat" e "kitten" – Figura 18.

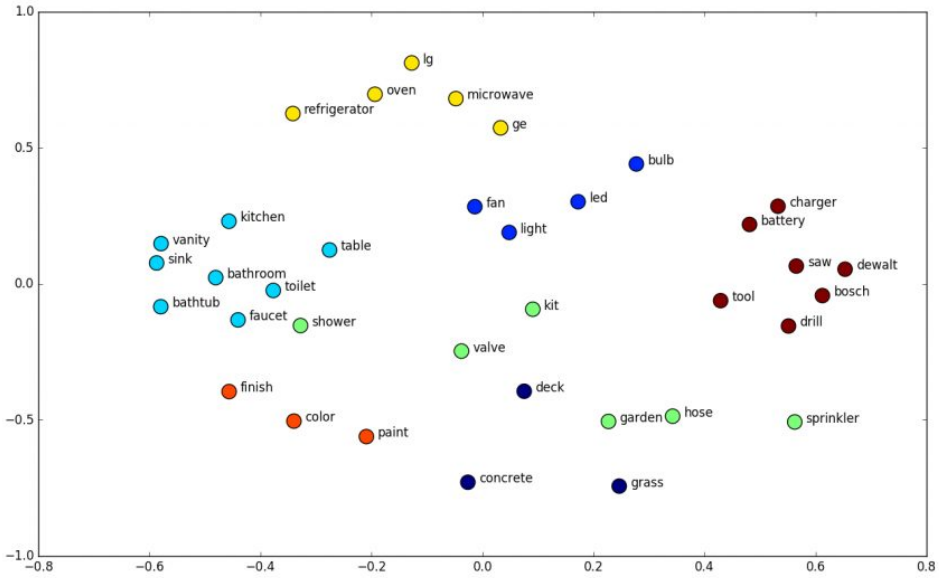


Figura 18 – Idealização da representação de palavras após tokenização e embedding: representação num espaço bidimensional – palavras semelhantes "ocupam" localizações próximas [75].

b) Codificação Posicional

Na arquitetura do Transformer não existe nativamente uma noção da ordem dos tokens. Para resolver esse problema, foi necessário adicionar um vetor de posição dentro de cada frase a cada *embedding* [73]. Este é composto por várias combinações de funções seno e co-seno, com valores que variam -1 a 1 (ver Figura 19). Por exemplo:

Posição 1 → `[0.1, 0.0, 0.1, 0.0, 0.1, 0.0]`

Posição 2 → `[0.0, 0.1, 0.0, 0.1, 0.0, 0.1]`

Posição 3 → `[0.1, 0.1, 0.0, 0.0, 0.1, 0.1]`

Mas este refere-se à posição de cada token/embedding na janela de contexto (*context window*), que nos LLMs mais recentes pode ter até 1 milhão de tokens. Assim, o vetor final de cada token é a soma: **Embedding + Posição** [73].

E vai ser este "novo vetor" que será processado – convertemos um conceito não numérico num elemento numérico condensando significado real ou abstrato, relações semânticas e posição dentro do nosso texto (janela de contexto do LLM utilizado) numa entidade numérica que pode ser processada matematicamente.

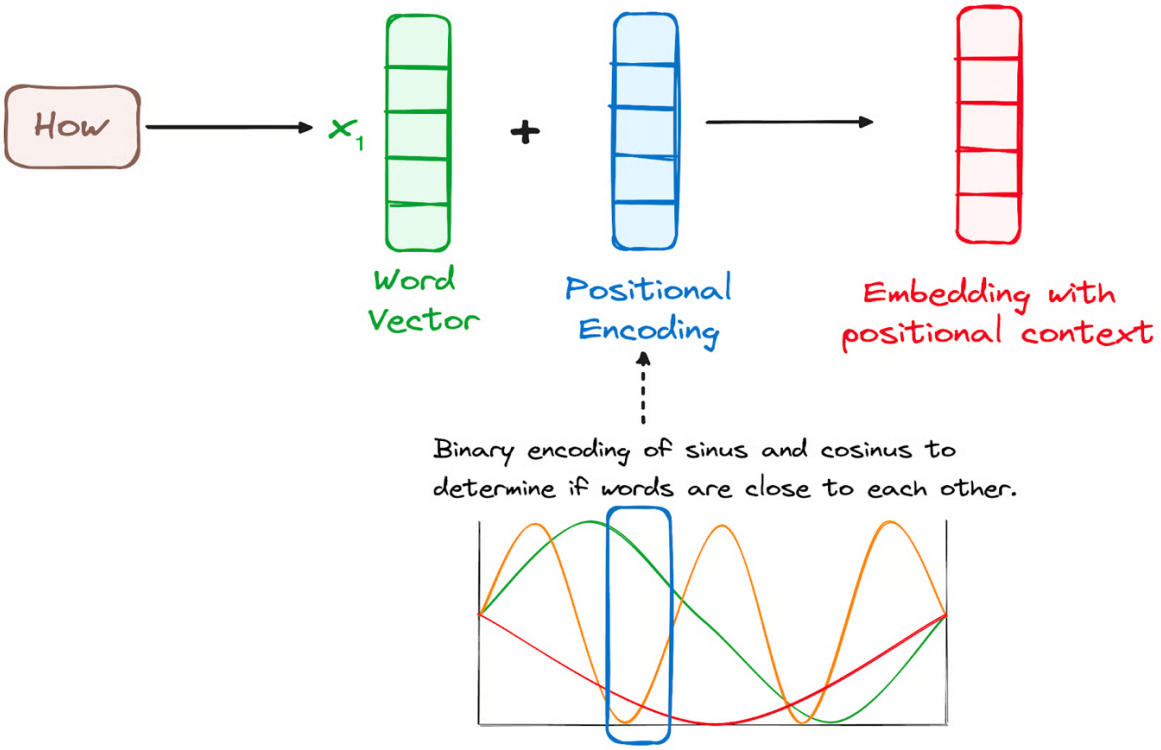


Figura 19 – Positional Encoding [73]

O número de dimensões (*embedding size*) de cada vetor neste espaço virtual depende do número de parâmetros do modelo:

- GPT 2 – 117 milhões de parâmetros – 768 dimensões
- GPT 3 – 175 mil milhões de parâmetros – 12 288 dimensões
- GPT 4 – sem resultados oficiais, mas estimado em 12 000 a 16 000 dimensões...
- Gemma 3 – 12 mil milhões de parâmetros – 4096 dimensões (um dos modelos utilizados no projeto desta Tese)

Ou seja, se quisermos adicionar mais informação em cada *token* (melhor caracterização → informação mais densa → maior granularidade) basta "apenas" aumentar o embedding size. Assim, a possibilidade de ter informação densa codificável matematicamente permite aos LLM uma aplicação ampla.

c) Mecanismo de Attention & Self-Attention (Q, K, V)

A "nova" arquitetura dos LLM derivou de dois conceitos fundamentais*: Attention e Self-Attention*. A atenção é um mecanismo que permite ao LLM "focar-se" em elementos mais importantes na janela de contexto, ou seja, as palavras / tokens deixam de ser todos "igualmente importantes", e passam a ter pesos diferenciadores (*scores*) [73]. Já o conceito de Self-Attention é mais complexo, exprimindo relações de longo alcance na janela de contexto.

Cada token é transformado em três novos vetores por multiplicação com matrizes de peso (aprendidas no treino):

Q (Query) → vetor que representa uma palavra / token da sequência de input no mecanismo de atenção; "de que tokens eu preciso? o que procuro?" (Questão)

K (Key) → vetor no mecanismo de atenção que corresponde a cada palavra ou token na sequência de input; "que informação eu trago? etiqueta visível para os outros tokens" (Rótulo)

V (Value) → conteúdo transmitido ou informação que pode ser passada.

Ou seja, quando o Query e a Key exibem elevada correspondência, esta traduz-se por um score de atenção elevado.

Voltemos ao exemplo simplificado (que geralmente codifica cada token em menos dimensões):

Para "Yellow" `[0.9, -0.2, 0.4, 0.1, -0.5, 0.6]` em "My Yellow Cat":

Q = `[0.2, 0.1, 0.7]`

K = `[0.3, 0.8, 0.2]`

V = `[0.9, 0.4, 0.6]`

A atenção calcula a relevância entre tokens, ou seja, cada "token atualizado" contém um pouco de todos os outros tokens**,** ponderado pela sua relevância na janela de contexto:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

- O produto QK^T mede a compatibilidade entre tokens (quantificando como a Query de um se alinha à Key do outro).
- O softmax converte em probabilidades (pesos normalizados).
- Multiplicar pelos Values devolve uma combinação ponderada da "informação" de cada *token.*

Isto cria uma matriz de pesos de atenção, mostrando quanto cada palavra "olha / se alinha" com as outras – ver Figura 20 e Figura 21. Exemplo: "Yellow" pode dar um peso elevado a "Cat", porque semanticamente estão ligados: cor ↔ objeto [73], [74], [76].

Mas cada conjunto Q / K / V pode reduzir a densidade de informação representada por cada *token*.

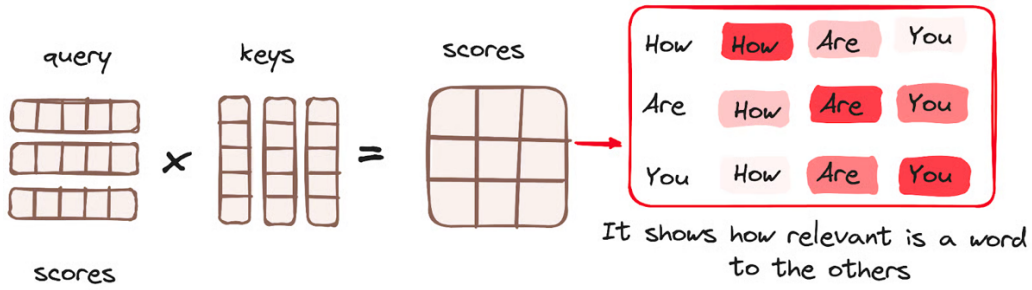


Figura 20 – Mecanismo de Atenção [73]

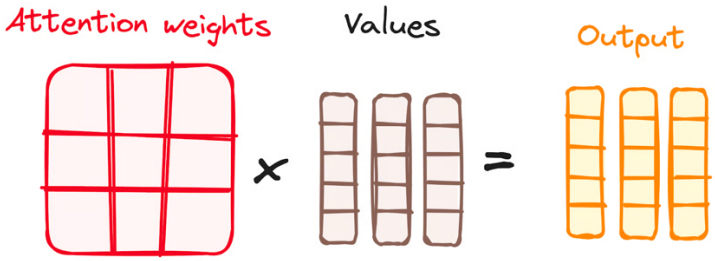


Figura 21 – Output final [73].

d) Multi-Head Attention

Em resposta ao problema anterior, este mecanismo permite "multiplicar" a análise em focos de atenção. Ou seja, em vez de um só mecanismo de atenção, o modelo cria várias Attention Heads em paralelo, aprofundando a análise na janela de contexto:

Attention Head #1 pode capturar sintaxe (ex.: sujeito-verbo).

Attention Head #2 pode capturar semântica (ex.: adjetivo-substantivo).

Os resultados são concatenados e projetados novamente para o espaço vetorial num "novo" vetor multidimensional (no nosso exemplo volta às 6 dimensões). E em cada Attention Head são gerados conjuntos de Q / K / V diferentes, mas que podem ser processados em paralelo. Ou seja, é possível "olhar" para todos os tokens na janela

de contexto não apenas sequencialmente, mas todos em conjunto, com relações múltiplas entre si e conseguir fazê-lo sob vários pontos de vista / análise [73].

e) *Feed-Forward Network and Layer Normalization* - Rede neuronal retropropagada normalizada

Após o mecanismo de atenção, cada vetor passa por camadas densas (*fully connected*) que permitem maior capacidade de abstração. Ou seja, cada "novo token pós mecanismo de atenção" vai ser processado individual e independentemente e submetido a transformação não-linear, podendo temporariamente passar das 6 dimensões do nosso exemplo para 36 e depois de novo para as 6 dimensões [76], [77], [78].

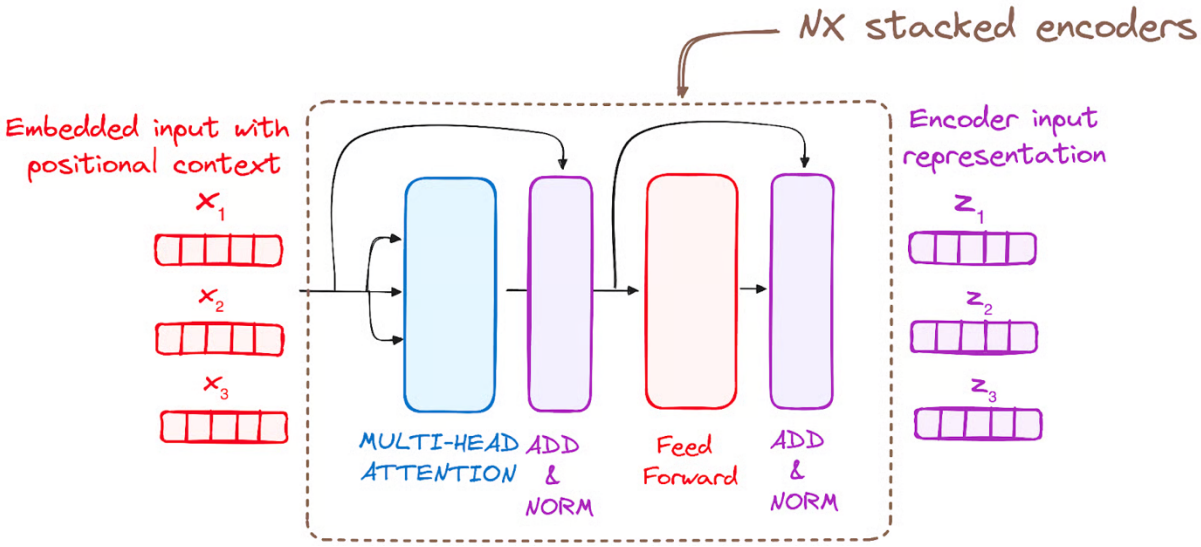


Figura 22 – Esquema global [73].

Ou seja, uma rede neuronal retropropagada transforma um input X num output Y, e de uma maneira geral teremos:

$$a_i^l = (w_i^l)^T h^{l-1} + b_i^l$$

$$h_i^l = f(a_i^l)$$

h^{l-1} → vetor de **saídas (ativação)** da camada anterior ou o inicial na 1ª camada.
w_i^l → vetor de pesos que liga a camada anterior ao neurónio i da camada l.
b_i^l → **bias** associado ao neurónio i da camada l, que é um parâmetro treinável.
a_i^l → soma ponderada das entradas antes da ativação (também chamado *pré-ativação*).
f → função de ativação (ReLU, tanh, sigmoid, GELU, etc.).
h_i^l → saída/ativação do neurónio i após aplicar f.

Este processamento pode ter alguns problemas:

- os valores de $\mathbf{h}_i^l$ podem crescer exponencialmente, com pesos muito elevados;
- diferentes tokens podem ter ativações muito diferentes ao conterem informação mais densa;
- uma camada pode gerar ativações muito elevadas e outra muito pequenas, acabando desalinhadas com escalas muito diferentes, gerando gradientes instáveis (*vanishing / exploding gradients);*
- janelas de contexto muito longas podem levar a grandes desequilíbrios de atenção.

Ou seja, a *Layer Normalization* – normalização de camadas - tem como papel condensar todas as ativações de um único token numa única camada e ajustar a distribuição dos valores das ativações: ajustando através da subtração da média; dividindo pelo desvio padrão e multiplicando pelo bias e pelo ganho.

No final, vai uniformizar as várias escalas possíveis, permitindo que cada token seja apresentado à camada seguinte num valor normalizado; vai manter a consistência entre tokens, permitindo que todos sejam tratados de forma equilibrada e, ao estabilizar os gradientes, evita explosão ou desaparecimento, tornando-o estável [77].

Voltemos ao exemplo "yellow":

$$a = [3.5, -1.2, 0.7, 5.1, -0.3, 2.0]$$
$$\text{Média } \mu = 1.6; \text{ Desvio padrão } \sigma \approx 2.2;$$
$$\text{Normalização: } a' \approx [1.0, -1.1, -0.3, 1.7, -0.7, 0.3]$$

Agora, todos os valores estão centrados e na mesma escala para serem apresentados à próxima camada. Se esta não for realizada, então o valor "5.1" dominaria a transformação; mas com a *Layer Normalization* "5.1" passou a ser "1.7 desvios acima da média da camada".

f) Estrutura do Descodificador *(Decoder)*

O Descodificador *(Decoder)* vai, essencialmente, criar sequências de texto com base numa entrada *(Input)*. A sua arquitetura é muito semelhante ao Codificador *(Encoder)*, exceto na camada Multi-Head Attention: aqui cada uma vai ter apenas uma função específica atribuída.

O processo inicia-se com o *Embedding do Input* (que é o *Output* criado pelo *Encoder*), a que por sua vez é também atribuído um *Positional Encoding*. Após isso, os vetores são alimentados no *Self-Attention*, semelhante ao encontrado no *Encoder*, mas com uma diferença: a relação dos tokens com outros tokens subsequente é bloqueada (*Masked Self-Attention*). Cada palavra na frase que está a ser gerada nunca pode ser influenciada por tokens futuros. De outra maneira, este Masked Self-Attention assegura que a probabilidade de um conjunto de tokens para a posição seguinte só pode depender dos conjuntos de probabilidade dos tokens anteriores.

No final, existe uma passagem por uma camada do Classificador Linear e uma camada SoftMax que transforma os valores anteriores em probabilidades (0-1), com as quais o modelo escolhe o token a apresentar (Figura 23).

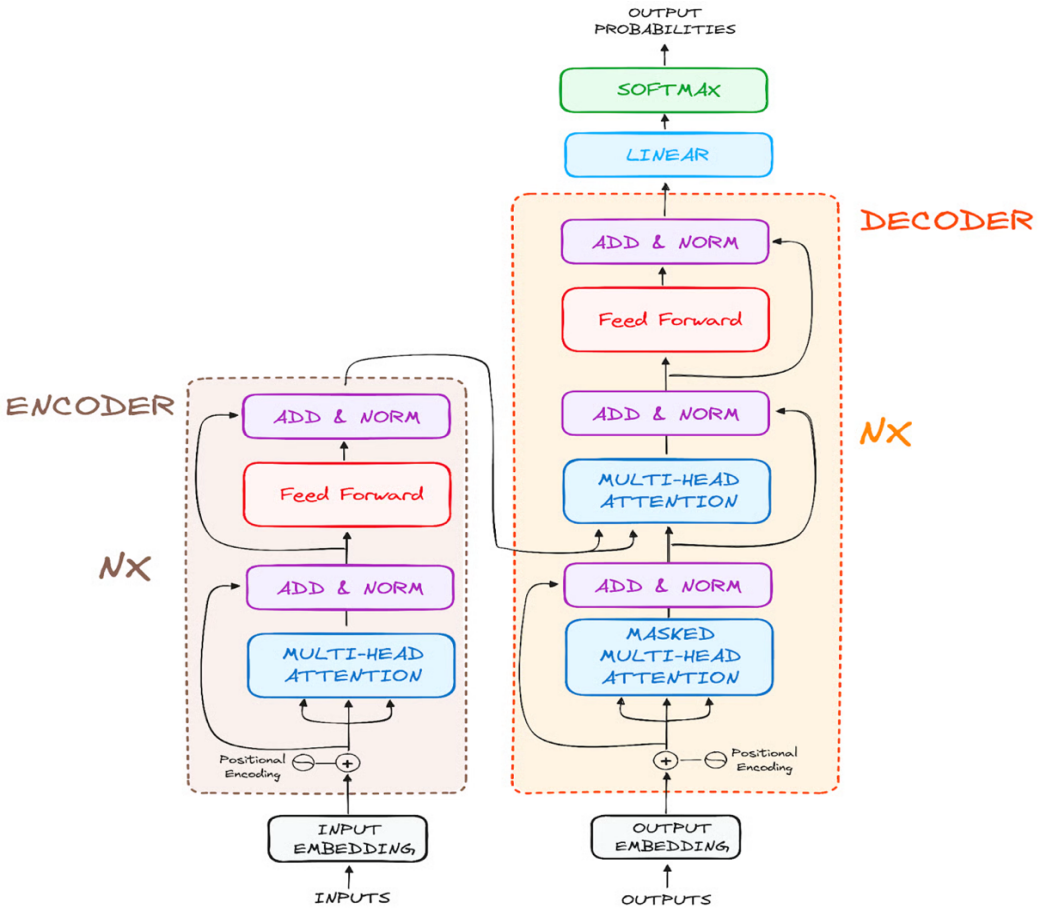


Figura 23 – Esquema final completo da arquitetura Codificador → Descodificador *(Encoder – Decoder)* [73]

5.2.2 Tipos de Arquitetura: Encoder-only, Decoder-only e Encoder-Decoder

a) Encoder-only

Em termos de estrutura apenas a parte Encoder do Transformer é utilizada; cada token é contextualizado através de self-attention bidirecional, ou seja, o modelo considera os tokens anteriores e posteriores em simultâneo. Utilizou no treino o *Masked Language Modeling* em que o modelo recebe frases com palavras mascaradas e deve prever os tokens escondidos. Consegue uma compreensão profunda do texto apresentado, mas não consegue gerar texto de forma auto-regressiva. Na área médica é utilizado para NER, classificação de documentos e extração estruturada de dados em textos médicos, se previamente treinado. Como exemplos temos o BERT©, BioBERT©, ClinicalBERT©**,** etc [79], [80].

b) Decoder-only

É a arquitetura mais usada em LLMs generalistas modernos.

Nesta arquitetura apenas a parte Decoder do Transformer é utilizada; é autoregressivo, isto é, prevê o próximo token a partir dos anteriores (atenção causal, que impede olhar para tokens futuros). O treino foi efetuado com recurso a *Causal Language Modeling (CLM)* — o modelo lê a sequência da esquerda para a direita e prevê o próximo token sequencialmente, permitindo a geração de texto fluida e contextual. Apesar de tudo, é menos eficiente em tarefas de compreensão pura sem *fine tuning* pelo facto de ser generalista. Tem sido utilizado em medicina para redação de relatórios clínicos, como Chatbots médicos ou assistentes de decisão clínica e apoio na elaboração de notas clínicas e resumos. Como exemplos temos: GPT-4©, LLaMA©, Gemma© [66], [67], [68], [69], [71], [72].

c) Encoder-Decoder

Esta arquitetura inclui Encoder e Decoder: Encoder processa o input completo e cria representações contextuais de cada token. Decoder gera a saída, usando self-attention (sobre tokens já gerados) e cross-attention (olhando para as representações do encoder). Pode ser excelente em tradução, sumarização e resposta a perguntas, mas é mais exigente do ponto de vista computacional porque utiliza dois blocos simultaneamente. O processo de treino utiliza tarefas de transformação *input* → *output*, formuladas em paradigma *seq2seq* (sequence-to-sequence). Neste modelo, o output pode não ter a mesma forma ou tamanho do input. Como exemplo pode resumir uma nota clínica em 2-3 frases mais importantes ou responder a uma questão sobre o texto processado anteriormente. Exemplos: T5©, BART©, SciFive©, BioT5© [78].

Resumidamente, o BERT destaca-se em tarefas de compreensão profunda (NER e / ou classificação), enquanto GPT é mais forte em geração fluida; o T5 elimina essa divisão ao utilizar o paradigma *text-to-text*. O BERT geralmente supera GPT quando *fine-tuned* em tarefas específicas, devido à sua natureza bidirecional de compreensão, mas falha se a tarefa for ligeiramente diferente da prevista; o GPT pode ser mais flexível; o T5 é o mais flexível dos três [70], [78].

A utilização de um LLM *decoder only* pode, até certo ponto, imitar um *encoder-decoder,* com vantagens na menor exigência computacional: a utilização de uma *prompt* bem desenhada delimita o que o *decoder only* pode fazer com o texto inserido na janela de contexto. Ou seja, podemos exemplificar: "segue um relatório clínico; devolve os valores dos campos [peso] e [idade gestacional]".

Mas uma elaboração correta da *prompt* é fundamental, como se aborda no ponto seguinte: 5.3 Prompts e Prompt Engineering.

5.3 Prompts e Prompt Engineering

A importância de uma prompt bem construída é de tal modo fundamental que motivou a publicação de um White Book sobre Prompt Engineering pela própria Google© [81].

Sempre que consideramos o input e output de um LLM, quase sempre pensamos num prompt de texto (por vezes acompanhado por outras modalidades, como prompts de imagem ou som ou vídeo), temos de compreender que este é o input que o modelo usa para prever um output específico. Apesar de qualquer pessoa poder escrever um prompt, para criar um prompt mais eficaz teremos um processo potencialmente mais complicado. Muitos aspetos do prompt afetam a sua eficácia: o modelo que utiliza, os dados de treino do modelo, as configurações do modelo, a escolha de palavras, estilo e tom, estrutura e contexto, tudo isso é importante. Portanto, a engenharia de prompts pode ser também um processo iterativo. Prompts inadequados podem levar a respostas ambíguas e imprecisas e podem prejudicar a capacidade do modelo de fornecer outputs significativos [81].

O LLM é, essencialmente, um motor de previsão: recebe inputs sob a forma de texto e prevê o próximo token, repetindo o ciclo e incorporando o que acabou de gerar no contexto para prever o seguinte. O *Prompt engineering* visa moldar essa sequência de previsões para obter o comportamento desejado.

Os parâmetros base que podemos definir no LLM são essencialmente os seguintes:

Número máximo de tokens na saída (*max tokens*)**:** define o número máximo de tokens que o modelo pode gerar até parar; não deve ser confundido como uma definição para ser mais conciso, apenas interrompe o fluxo de geração; mais tokens / limite mais elevado implica maior custo computacional.

Temperatura: controla o grau de aleatoriedade na escolha do *next token*; quando assume o valor mínimo (0) o modelo escolhe sempre o mais provável e os resultados são mais estáveis; se o objetivo for a obtenção de resultados mais criativos, então deve ter um valor mais elevado (>1), mas aumenta o risco de incoerência e resultados absurdos. Ou seja, para temperaturas máximas, a probabilidade de todos os tokens é igual – distribuição plana [82].

Top-K: globalmente, o modelo prevê uma distribuição de probabilidades para todos os tokens existentes no modelo; neste caso, podemos limitar este conjunto a um número (K), ou seja, os K mais prováveis. A escolha é depois realizada apenas dentro deste conjunto de K tokens. Valores mais baixos permitem resultados mais coerentes [82].

Top-P: Aqui podemos definir um limite inferior de probabilidade acumulada; ou seja, o número de *tokens* considerados, tal como no Top-K, é definido através da soma das suas probabilidades individuais, sendo que o número de elementos do conjunto é variável. pode ser um parâmetro de segurança importante: se um dos *tokens* considerado já possuir uma probabilidade igual ou superior à definida (P),

então apenas esse vai ser considerado, impedindo a dispersão do modelo por *tokens* irrelevantes, o que não ocorre no Top-K. Também permite considerar a "incerteza" – pois obriga a considerar múltiplos *tokens* quando as probabilidades são relativamente baixas, isto é, não existe um número relativamente limitado de *tokens* para escolher, como se mostra na Figura 24 [82].

Genericamente, a maioria dos LLMs aplica inicialmente o Top-K e depois aplica o Top-P. Mas se a temperatura for definida como "0", então P = 1 e K = 1...

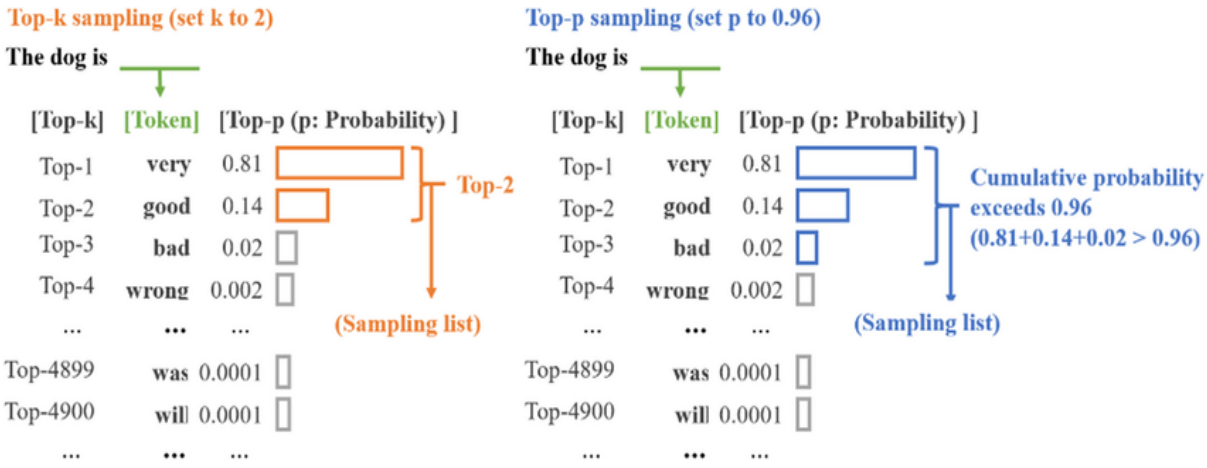


Figura 24 – Processo de amostragem Top-P e Top-K [82].

Técnicas de Prompting:

Zero-Shot: um texto simples e direto, em linguagem natural; sem limites e sem exemplos. Pode ser falível. Funciona melhor em tarefas simples de classificação desde que temperatura seja ajustada para valores próximos de "0" [81].

One-Shot e Few-Shot: aqui pode ser fornecido um ou mais exemplos do que pretendemos; ou seja, o modelo pode tentar replicar o que foi fornecido. Se possível, devem ser obtidos exemplos que ilustrem "limites" – *edge cases*. Ou seja, ao apresentar mais do que apenas os casos "normais", estamos a definir também os limites do possível, evitando respostas erradas ou alucinação quando o modelo vai processar casos com dados mais raros. Mas não devem ser apresentados múltiplos casos, apenas os mais emblemáticos (normal e extremos admissíveis) [81].

System, Contextual and Role Prompting: define o comportamento do modelo; no system prompting é orientado o tipo de tarefa principal (tradução, pesquisa, elaboração de texto criativo, etc.). No contextual prompting é fornecido o contexto

em que a tarefa vai ser desenvolvida – exemplo: baseia a resposta nos guidelines X; deve ser específico. No role prompting definimos o papel que pretendemos que o LLM assuma (se algum) – pode atuar como médico ao analisar os guidelines X ou como guia turístico ou ainda como professor, por exemplo [81].

Step-Back Prompting: A tarefa é subdividida em duas fases; na primeira fase é pedida uma revisão geral global de um determinado tema (ex. quais os tipos de ventilação existentes em neonatologia?) e numa segunda fase apresenta-se o verdadeiro problema (ex. qual a melhor estratégia de ventilação num extremo prematuro com hipoplasia pulmonar?). Ou seja, o primeiro prompt dá o contexto e ativa o conhecimento em background; habitualmente gera respostas mais completas e assertivas ao segundo prompt [81].

Chain of Thought (CoT): o objetivo é desdobrar o processo utilizado para solucionar um problema nos seus passos intermédios. Este tipo de estratégia permite resolver problemas mais complexos com modelos mais "modestos"; a tarefa está decomposta em passos mais simples. Também é possível aprender com a solução elaborada pelo modelo e, eventualmente, identificar o erro se este ocorrer. Este tipo de prompt também permite obter soluções mais reprodutíveis entre vários modelos diferentes; ou seja, menos permeável a diferenças no comportamento. Também permite prescindir de fine-tuning. Obviamente que produz mais tokens para chegar a uma determinada solução, o que pode aumentar o "custo" (energético e/ou económico) [81].

Self-Consistency: Esta estratégia permite melhorar a hipótese de uma resposta correta. Ou seja, a mesma prompt é apresentada ao mesmo modelo, sequencialmente, várias vezes. Se a temperatura >0, podemos admitir que várias respostas podem ser apresentadas. Posteriormente, o modelo vai utilizar as diferentes respostas, e gerar uma nova distribuição de probabilidade, apresentando a mais elevada como resposta final [81].

Tree of Thoughts (ToT): integra as duas últimas estratégias. Usa uma definição mais alargada de Chain of Thoughts ao permitir que o modelo explore diferentes cadeias de raciocínio "em simultâneo" em vez de uma única sequência linear (Figura 25). Pode ter vantagens ao explorar soluções menos evidentes em tarefas complexas [83].

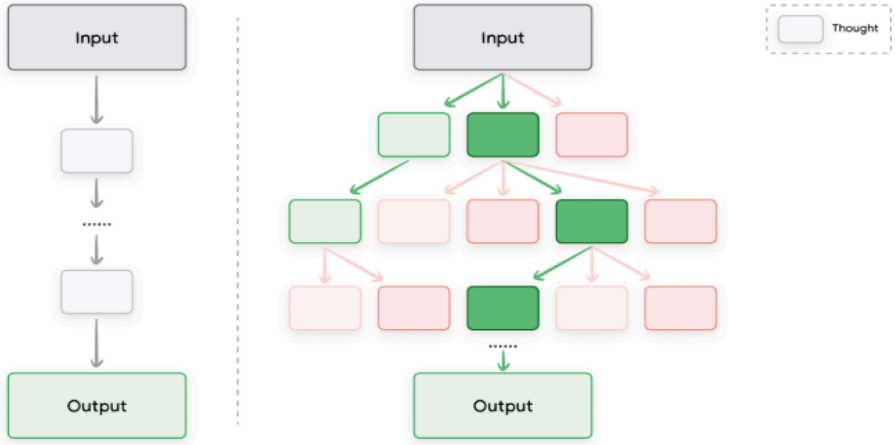


Figura 25 – Tree of Thoughts (ToT) [81].

Reason and Act (ReAct): esta estratégia tenta replicar o modo como o cérebro humano opera no mundo real. Permite alternar raciocínio em linguagem natural e ações externas como pesquisa ou cálculo; cria um ciclo:

pensamento → ação → observação

no final, os dados obtidos pela observação são inseridos no novo ciclo de pensamento → ação → observação. Estes terminam quando é encontrada uma solução final [81].

→ **ReAct vs Agentic AI:**

ReAct – o utilizador controla e supervisiona o ciclo pensamento → ação → observação, ou seja, como este deve ser orientado. É uma técnica de prompting que pode estar limitada à *context window* atual (Figura 26).

Agentic AI – o próprio modelo define e supervisiona o(s) ciclo(s); implica memória de curto e longo prazo; planeamento e objetivos intermédios; pode utilizar mais ferramentas e, eventualmente, alguma atitude crítica antes de apresentar a solução final. Constitui uma arquitetura de sistema que permite definir todo um plano de sub-tarefas [84].

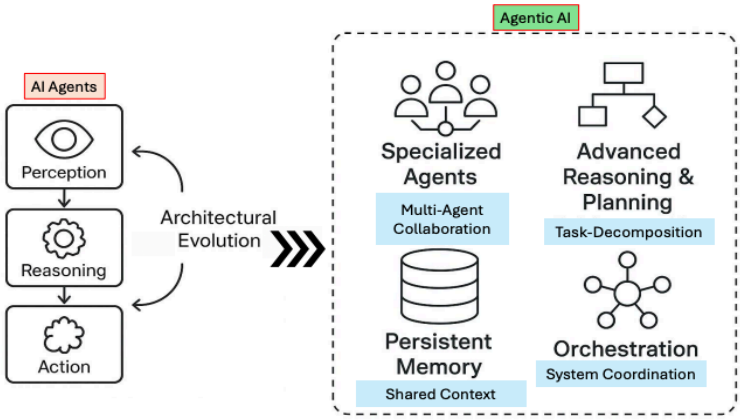


Figura 26 – Comparação entre o LLM como "Agente" (ReAct prompting) e Agentic AI [84].

Automatic Prompt Engineering (APE): atendendo à complexidade para resolver alguns problemas, pode ser útil usar o próprio LLM para criar várias prompts e escolher a(s) mais eficaz(es). No exemplo utilizado nesta referência, o LLM criava 10 prompts diferentes para treinar um chatbot para interagir com clientes, ou seja, 10 maneiras diferentes que os clientes poderiam utilizar para pedir a mesma coisa [81].

5.4 Optimização, Fine-Tuning, Ajustes

Até ao momento, abordámos os ajustes possíveis de LLMs sem alterar pesos:

- Prompt Engineering
- Sampling (Top-P, Top-K, Temperatura, MaxTokens)

Mas existem outras intervenções possíveis para optimizar o LLM.

5.4.1 RAG – Retrieval Augmented Generation

Não é um ajuste, não modifica o modelo. Também não é uma opção dentro do prompt engineering. Acaba por ser apenas uma adaptação leve ou mais propriamente um complemento. Pode ser definida como uma arquitetura para otimizar o desempenho de um modelo, conectando-o a bases de conhecimento externas [85].

Os LLMs são treinados em grandes conjuntos de dados e consultam essas informações para gerar resultados. No entanto, os conjuntos de dados de treino são finitos e limitados à informação à qual teve acesso — *corpora* de domínio público, artigos e outros dados acessíveis ao público.

O RAG permite o acesso a bases de conhecimento externas adicionais, como dados organizacionais internos, revistas académicas e conjuntos de dados especializados. Ao integrar informação relevante no processo de geração, os LLMs podem criar conteúdos específicos do domínio mais precisos, sem necessidade de treino adicional [85]. O esquema geral de funcionamento está tipificado na Figura 27.

Como exemplo, o LLM utilizado neste projeto (Gemma 3), funcionando offline, apenas é capaz de responder a questões sobre factos ocorridos até 2023 (dataset atualizado à data de treino; os parâmetros ficaram congelados).

Esta abordagem contorna as principais limitações dos modelos tradicionais, tais como a sua incapacidade de aceder a conhecimento em tempo real ou específico de um domínio, e diminui o erro quando analisa prompts com entidades fora do vocabulário ou raras [85].

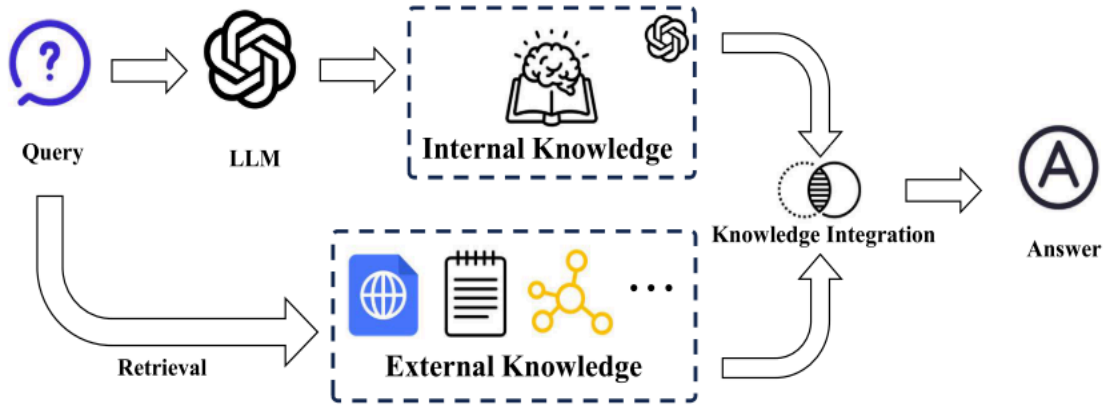


Figura 27 – Visão global do processo RAG [85].

Para obter resultados funcionais, o processo inerente ao RAG implica:

i. Precise User Intent Understanding – vários utilizadores diferentes vão utilizar prompts diferentes, mesmo que o objetivo seja, essencialmente, igual. O modelo deve compreender estas diferenças e ajustar a sua resposta ao utilizador.
ii. Accurate Knowledge Retrieval – é a fase mais crítica e influencia diretamente a qualidade do Output. O modelo deve aceder ao conhecimento de que necessita na forma mais atualizada e relevante, mas sem ficar "preso" numa pesquisa sem fim.
iii. Seamless Knowledge Integration – implica um alinhamento quase perfeito entre o conhecimento interno do modelo e o formato do conhecimento externo a ser adicionado. Deve também evitar contradições entre o interno e o externo.
iv. Superior Answer Generation – a resposta do LLM deve integrar todas estas informações, com estratégias de validação e referenciação cruzada para manter a exatidão da informação do Output. A utilização de Knowledge Citation é também fundamental.
v. Comprehensive RAG Evaluation – este processo integra duas atividades – busca e geração; ambas devem ser avaliadas em conjunto, e continuamente adaptadas ao utilizador.

O treino de modelos RAG envolve equilibrar a otimização quer da tarefa de busca de informação, quer a tarefa de geração de texto.

Estes modelos foram também adaptados a funcionar como Multimodal RAG, englobando conhecimento não apenas em texto, mas também imagem, som, vídeo [85].

Mas também podemos modificar mais profundamente os parâmetros do modelo, com *fine-tuning*.

5.4.2 Fine-Tuning Supervisionado, Instruction Tuning, Fine-Tuning específico de domínio, Multi-task Fine-Tuning e Chain of Thought Supervisionado

Instruction Tuning ou Supervised Fine Tuning refere-se ao processo de treino adicional de LLMs com um conjunto de dados que consiste em pares (instruction - output) de forma supervisionada, permitindo preencher a lacuna entre o objetivo de previsão do next token nos LLMs e o objetivo do utilizador, ou seja, que o LLM consiga seguir as instruções dadas [86].

Isto é frequentemente realizado a partir de exemplos adicionais que representam a forma como queremos que ele se comporte em aplicações específicas. Os LLMs são treinados em extensos corpora de textos para que possam realizar a previsão do next token; mas estes podem não ter um contexto que permita a correta compreensão das instruções do utilizador (Figura 28).

No limite, imaginemos que o utilizador pretende que o modelo extraia dados de um texto clínico e os apresente num formato universal (JSON, CSV, etc.) mas o modelo devolve os dados corretamente em linguagem natural, porque não tem conhecimento sobre como construir os formatos pedidos.

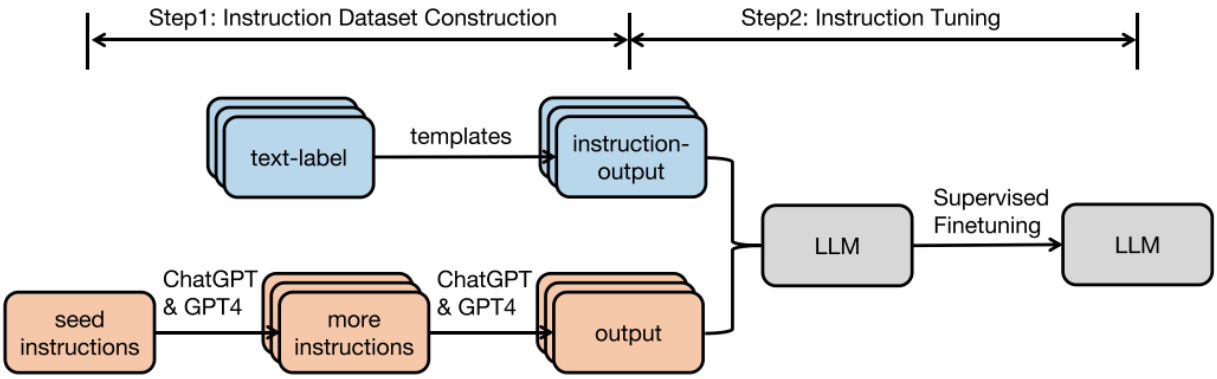


Figura 28 – Processo exemplificativo de *Instruction Tuning* [86].

Ao ser treinado posteriormente com datasets de pares instruction - output específicos, o processo é muito mais rápido pois parte do background já existente, e não implica um treino de um modelo em branco ou a modificação de arquitetura destes.

Apesar das vantagens, não está isento de problemas: a idealização destes datasets não é trivial e os exemplos tendem a ser limitados em qualidade, quantidade e criatividade; pode ajustar o modelo apenas nas instruções apresentadas como exemplo e a generalização ser mais difícil; e ainda alguma preocupação que o modelo apenas modifique superficialmente e não compreenda em profundidade as instruções recebidas [86].

Estes *datasets* podem, genericamente, ser divididos em:

1. Human-crafted Data – Natural Instructions© P3 Public Pool of Prompts©, XP3 Cross Lingual Public Pool of Prompts©, FLAN 2021©, LIMA ©, SuperNatural Instructions©, Dolly©, Open Assistant Conversations©;

2. Synthetic Data via Distillation – Alpaca©, WizardLM©, Evol-Instruct©, ORCA©, Baize©, etc.

3. Synthetic Data via Self-improvement – o modelo cria o output ou a instruction de acordo com o que falta e integra-as em si mesmo.

4. Reasoning Datasets – PRM800K©, O1Journey©, MathGenie©, DeepSeekMath© [86].

No final, consegue-se que um modelo mais modesto consiga ter um desempenho equivalente a modelos mais complexos. Um exemplo: o modelo Alpaca© 7B[2] resulta do fine-tuning do LLaMA© 7B com um dataset gerado no InstructGPT© 173B e consegue ter um desempenho próximo do InstructGPT© 173B, apesar do número muito inferior de parâmetros. Assim, temos um modelo com elevado desempenho num determinado domínio, preservando o core central do LLM original.

Apresenta alguns riscos: Overfitting quando o dataset é pequeno e a generalização é mais difícil; Catastrophic forgetting quando o modelo integra um novo conjunto de conhecimento e esquece partes do anterior.

Existe ainda a possibilidade de ajustar o modelo propondo a realização de tarefas combinadas (ex. em neonatologia – classificação + extração + sumarização), aumentando a robustez do modelo. Os datasets devem manter a proporção entre as várias tarefas pedidas.

No caso de se pretender um modelo mais dedicado ao raciocínio, pode ser utilizado um dataset com raciocínio explícito, i.e. decompondo os passos intermédios - Chain of Thought (CoT) Supervisionado. O modelo aprende como deve proceder para chegar a um determinado resultado, melhorando significativamente os resultados apresentados [81].

5.4.3 Multi-modality Fine-Tuning

Apesar dos LLMs estarem, sobretudo, vocacionados para processar linguagem escrita / texto (NLPs), e o input-output ser texto, podem fazer mais. Nesta modalidade, o modelo é ajustado para lidar com múltiplos tipos de dados além do texto: imagens, áudio, vídeo, sinais raw de dispositivos médicos, etc. Atualmente constitui a base para a análise de imagem em medicina e sua interpretação / relatório [86], [87].

[2] **B** – Notação original *Billion* – Milhares de Milhões. Optou-se por manter a notação original por ser a mais universalmente utilizada

Resumidamente, a arquitetura do modelo é modificada para aceitar um Input Multimodal: cada modalidade terá um encoder próprio adequado (transformer para texto e CNN para imagem, por exemplo). Seguidamente, os encoders vão converter tudo isto em vetores num espaço comum, sendo que o LLM "aprende" a fundir essas representações diferentes.

Depois, há que repetir os processos referidos (ver 5.4) com um dataset preparado para Supervised Multimodal Instruction Tuning (ex. Imagem: radiografia de torax neonatal + Instrução (texto): qual o diagnóstico? → Output (texto): compatível com pneumonia lobo superior direito).

Obtemos assim um MLLM – Multimodal LLM, adaptado a Med-VQA (medical visual question answering) e MRG (medical report generation), como o Med-Flamingo© e o LLaVa-Med©. A arquitetura final possível apresenta-se na Figura 29 – Arquitetura de um MLLM [87].

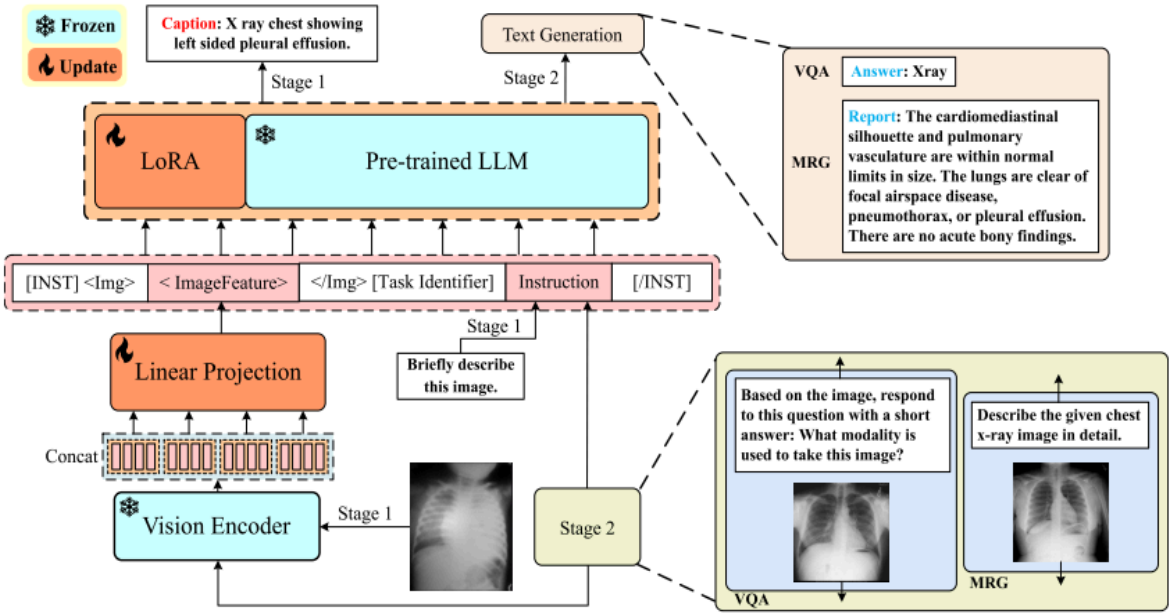


Figura 29 – Arquitetura de um MLLM [87]

5.4.4 Parameter-Efficient Fine-Tuning

Atualmente a questão dos ajustes centra-se cada vez na eficiência, o PEFT (Parameter-Efficient Fine-Tuning). O objetivo é ajustar apenas uma pequena fração dos parâmetros, mantendo a maioria congelada, evitando re-treinar todo o modelo. Para além de ser mais rápido e utilizar menos recursos, reduz o risco do *Catastrophic Forgetting*. Além disso, é escalável e adaptável, ou seja, pode-se adequar um MLLM à área da pediatria e, com a mesma base, adequar à neonatologia [88].

Este PEFT pode ser conseguido recorrendo a:

i. Adapters – constitui uma camada intermédia entre as várias camadas dos transformers; apenas esta é modificada / treinada, mantendo as restantes (originais) intocadas. Os restantes parâmetros do LLM ficam inalterados.
ii. Low Rank Adaptation (LoRA) – em cada transformer, a matriz original é complementada por outra de menores dimensões; reduz drasticamente o número de parâmetros que são modificados pelo treino; estes módulos podem ainda ser ligados/desligados consoante a necessidade. Existem também algumas variedades como o DyLoRA, AdaLoRA, IncreLoRA, etc.
iii. Prefix Tuning – implica a utilização de um prefixo, um conjunto de vetores extra, que são adicionados a cada camada do *transformer*. Resumidamente, no mecanismo de *Attention*, cada *token* "olha" para os anteriores; com o *Prefix Tuning*, são introduzidos tokens artificiais pré-treinados / adaptados antes dos *tokens* reais. Ou seja, estamos a modificar a linha de previsão de *tokens*. É uma maneira simples e leve de modificar um modelo, mas também a menos potente.
iv. Soft-Prompt – para além da prompt fornecida pelo utilizador, o modelo já utiliza uma outra prompt, muitas vezes invisível ao utilizador, que ajusta o Output. Existem múltiplas variedades (*late prompt tuning; instance dependent tuning, decomposed prompt tuning*, etc.). Exemplo: peço frequentemente os valores limite de uma determinada métrica; após algumas repetições para converter os valores em Unidades SI, o modelo passa a apresentar, logo de início, os resultados com Unidades SI.

Podem ainda ser combinados diferentes métodos de PEFT num Hybrid PEFT [88].

5.5 Problemas Inerentes aos LLM

5.5.1 Alucinação

A alucinação nos LLMs constitui, provavelmente, o mais comum e mais complexo problema, comprometendo severamente a sua fiabilidade. Para além disso, este problema persiste, apesar de todos os meses surgirem novos avanços nos modelos online, *state of the art,* as verdadeiras montras tecnológicas com investimentos avultados.

A alucinação pode ter consequências significativas para aplicações no mundo real. Por exemplo, um modelo de IA na área da saúde pode identificar incorretamente uma lesão cutânea benigna como maligna, levando a intervenções médicas desnecessárias. As alucinações também podem contribuir para a disseminação de informações erradas, sem prévio *fact check* [89].

Então a questão é: será um problema inerente aos LLM e IA? Estará integrado na sua estrutura? Será um problema de programação? Recentemente a própria OpenAI© (GPT) abordou este problema [90].

O que é uma alucinação?

No DSM-5[3], uma alucinação é definida como uma perceção falsa que ocorre na ausência de um estímulo externo real do órgão sensorial relevante, ou seja, uma "perceção sem objeto". É uma manifestação de psicose, na qual o indivíduo percebe algo que não está presente no mundo exterior, como ouvir vozes ou ter visões.

No caso da IA, a alucinação pode ser definida como um fenómeno em que, num modelo – frequentemente um chatbot de IA generativa ou uma ferramenta de visão computacional – identifica padrões ou objetos que são inexistentes ou impercetíveis para observadores humanos, criando resultados sem sentido ou totalmente imprecisos. Contrariamente ao fenómeno humano, aqui é um fenómeno "estatístico".

Mas será a alucinação sempre um problema? Será a criatividade "alternativa" útil? Certamente que sim [89]:

1. Arte e Design – a IA consegue criar imagens com conceitos diferentes e inovadores, surreais possivelmente, mas sempre com criatividade.
2. Análise e Visualização de Dados – ao evidenciar conexões inesperadas e assim definir tendências antes não visualizáveis em processos tradicionais
3. *Gaming* e Realidade Virtual – a possibilidade de criar alucinações e gerar ambientes virtuais pode ajudar os desenvolvedores de jogos e designers de

[3] **DSM-5**: Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5), 2013, American Psychiatric Association.

realidade virtual a imaginar novos mundos que elevam a experiência do utilizador a um novo patamar. As alucinações também podem adicionar um elemento de surpresa, imprevisibilidade e novidade às experiências de jogo.

Então qual a génese deste fenómeno? Podemos genericamente dizer que o *overfitting,* a qualidade dos dados e respetivo *bias,* e a complexidade dos modelos são as mais frequentemente identificadas.

Segundo a OpenAI [90], na análise que faz de modelos de elevada complexidade, identifica ainda outras causas-raíz deste problema, distribuídas pelas fases de pré-treino e pós-treino:

a. Fase de Pré-Treino - Erros Estatísticos

O pré-treino dos LLMs é feito com um objetivo simples: prever o próximo token numa sequência de texto. Este objetivo gera inevitavelmente alucinações (saídas incorretas, mas plausíveis), mesmo que o texto seja perfeito.

Alguns dados fornecidos ao modelo na fase de aprendizagem são factos raros e infrequentes, ou seja, um problema de *long-tail samples*: os dados seguem distribuições com muitos eventos comuns e raros, logo o erro em eventos raros é quase inevitável.

Se um modelo for relativamente pouco complexo, pode não conseguir distinguir co-ocorrências de verdadeiras relações, ficando por representações superficiais – *spurious correlations***.**

Pode ocorrer também *exposure bias*: durante o treino o modelo só vê sequências corretas; na fase de (auto)geração de texto, tem de lidar com as suas próprias predições, o que pode criar erros em cascata.

A *epistemic uncertainty* resulta de não existir conhecimento suficiente para responder corretamente; ou seja, o corpus de treino não contém a informação necessária ou contém muito poucos exemplos de determinada questão: corresponde a um vazio de conhecimento.

GIGO – Garbage In, Garbage Out: neste caso, os dados existem, mas estão incorretos (fake news, informação desatualizada, dados cientificamente incorretos). O modelo aprendeu esses dados / "erros". Há um erro herdado e propagado.

Em todas estas causas-raíz o cenário é semelhante: a opção do modelo devia ser *IDK – I Don't Know,* mas como é suposto gerar sempre um token seguinte, acaba a selecionar a opção mais plausível (estatisticamente). Mais exatamente, cada token possível é escolhido após o modelo calcular uma distribuição de probabilidades para o próximo token [90].

Aqui teremos em ação os conceitos de "temperatura", "Top-K" e "Top-p", permitindo maior ou menor criatividade e aleatoriedade na escolha do token seguinte – aleatoriedade controlada – *decoding randomness***.**

Efeito Bola de Neve – *snowballing effect:* Cada token gerado torna-se o input para prever o seguinte. Se logo no início é escolhido um token "incorreto[4]", todos os próximos cálculos de probabilidade vão basear-se nesse elemento, apesar de perfeitamente coerentes com essa opção.

b. Fase de Pós-Treino - Persistência

Nesta fase há um ajuste (*fine-tuning*) do modelo, só que as estratégias até hoje podem, inadvertidamente, incentivar a alucinação em vez de a reduzir [90]. Um modelo antes de ser lançado é avaliado em vários *benchmarks*:

MMLU (Massive Multitask Language Understanding) - conjunto de 57 disciplinas diferentes (ciências, humanidades, medicina, direito, etc.), com mais de 15.000 questões de múltipla escolha. O modelo deve escolher sempre uma opção, nunca podendo referir a opção que assume a incerteza (IDK). A pontuação é binária (0 se resposta errada, 1 se resposta certa).

GPQA (Graduate-Level Google-Proof Q&A) – conjunto de perguntas de nível pós-graduado (ciências naturais e matemática), que não pode ser resolvido apenas com pesquisas ao nível do Google, testando raciocínio avançado. Mais uma vez, classificação binária 0-1.

SWE-bench (Software Engineering Benchmark) – programação e engenharia de software; avalia se o modelo consegue resolver bugs e implementar patches em repositórios reais de código (extraídos do GitHub). A classificação é obtida pela análise da resposta proposta – saída de código.

HELM (Holistic Evaluation of Language Models) - Framework criado por Stanford para avaliar LLMs de forma multidimensional. Avalia exatidão, fidedignidade, viés, eficiência, toxicidade. Conjunto de cenários (sumarização, Q&A, tradução), com métricas diferentes em cada um, ou seja, um painel de avaliação. Ainda assim, muitas tarefas têm pontuação binária (correto/incorreto – 1/0). Nem aqui se facilita a opção IDK.

Ou seja, em nenhum deles se promove a opção de incerteza (IDK), podendo assumir, de uma forma simplista, que o modelo é incentivado a arriscar sempre uma opção, sendo que a avaliação é sempre binária (0-1), só são incentivados os "1" e o IDK = 0.

Após a fase de benchmarking, o modelo pode ser ajustado através do *RLHF - Reinforcement Learning with Human Feedback*, em que humanos avaliam várias respostas do modelo e escolhem a "melhor", só que clareza, confiança, estilo, detalhe podem estar erradas. Respostas mais longas, estruturadas e com tom de autoridade tendem a receber melhor pontuação, mesmo que factualmente erradas: parecer certo versus

[4] o termo "incorreto" é também incorreto, pois o modelo não sabe se o que está a escolher é verdadeiro ou falso, apenas sabe a distribuição do mesmo nos dados de pré-treino e a probabilidade inerente a essa opção como próximo token.

estar certo. Ao selecionar respostas que parecem corretas, mas não o são, temos o conceito de *misleading alignment training* [90].

Para além destes, mais recentemente surgiram dois fenómenos: *reversal course e context hijacking*. No primeiro caso, o modelo tem dificuldade em inverter um raciocínio simétrico (Paris é a capital de França. → Qual é a capital de França? → Londres...). No segundo caso, o modelo pode ser enganado pelo próprio contexto que lhe é dado no prompt ou via RAG, quando disponível. Muitos modelos utilizam o RAG acedendo a mais fontes para procurar uma resposta que não está na base de dados/parâmetros do LLM. Só que a consulta de bases exteriores aos parâmetros do LLM pode também utilizar dados errados ou enviesados.

De volta ao nosso exemplo:

Imaginemos que o nosso LLM devia continuar a sequência "*My" "Yellow" "Cat*" e em vez de optar pela sequência *"is" "sleeping" "on" "the" "sofa",* as opções e probabilidades previstas são:

"is" → 0.60; "was" → 0.25; "ate" → 0.15 e a opção escolhida não foi a de maior probabilidade mas sim a de menor - "ate" - pela aleatoriedade (temperatura, top k e top p).

Sequencialmente, teríamos:

"food" → 0.50; "a mouse" → 0.30; "the sofa" → 0.20 e a opção escolhida foi "the sofa".

Assim, temos presentes várias das causas referidas anteriormente para uma alucinação. A frase é gramatical e estruturalmente correta, mas absurda. Resta saber se o pré-treino do modelo incluiu a informação de que os sofás não são comestíveis ou a utilização de RAG incluiu Comics da Disney...

Em resumo, de acordo com o trabalho mais recente, a alucinação em modelos de IA, especialmente LLMs, é estatisticamente inevitável (pré-treino) e reforçada por incentivos errados (pós-treino) [89], [90].

Nomeadamente, prova-se que se o modelo tiver um erro de 10% numa tarefa de classificação (verdadeiro/falso), a probabilidade de alucinação quando gerar texto utilizando esse parâmetro/dado será duas vezes superior (20%) [90].

Estratégias para minimizar a alucinação em LLMs:

Do acima exposto se conclui que não é por incluir mais dados / mais parâmetros e mesmo utilizar o RAG que a alucinação pode ser evitada. Globalmente, a questão pode ser colocada acerca do modo como se constroem os *benchmarks*; a incerteza deveria ser diferenciada da resposta errada; o erro deve ser fortemente penalizado, mas a incerteza ou ignorância deve ser possibilitada e não colocada no mesmo nível [89], [90].

Estes devem passar a incluir a opção "IDK" e atribuir pontos; paralelamente, podem também ser adicionados "intervalos de confiança" na resposta dada (apresentando duas ou três opções possíveis e respetivas probabilidades / grau de certeza).

Igualmente, a definição de um grau de certeza mais baixo permite diferenciar alternativas entre as várias opções; apenas se existir uma única opção com elevado (acima do limiar mínimo) grau de certeza, então será apresentada apenas uma resposta. Isto pode ser conseguido por *behavioral calibration.* O modelo também pode aprender a expressar a confiabilidade da resposta (a opção é Paris com 88% de certeza) [90].

5.5.2 Viés e Equidade

O uso de LLMs também levanta preocupações em relação ao potencial preconceito.

Os LLMs aprendem com os dados com os quais são treinados, e esses dados podem refletir preconceitos sociais existentes, incluindo aqueles relacionados a raça, etnia, género, status socioeconómico, idade, orientação sexual ou deficiência. Quando esses preconceitos são codificados nos LLMs, podem ser amplificados e perpetuados, levando a consequências prejudiciais, sobretudo no domínio sensível da Saúde [91], [92].

Esse tipo de viés surge quando os dados de treino não refletem com precisão a diversidade real das populações, apresentações clínicas ou fatores relevantes. Vários estudos destacaram os efeitos prejudiciais de conjuntos de dados distorcidos: por exemplo, desequilíbrios na representação de tons de pele em datasets de imagens dermatológicas. Estudos que analisaram a legibilidade e o conteúdo do texto gerado por LLM descobriram que os resultados variavam significativamente com base na raça e status socioeconómico. Além disso, a utilização de dados históricos pode também refletir preconceitos sociais do passado, mesmo que tecnicamente "precisos"; e conduzir os LLMs a reproduzir e amplificar esses preconceitos nos resultados. Alguns estudos referem prevalências de 93% para viés ligado ao género e 90% relacionado com etnias [92].

Para além disto, é necessário referir que cada modelo / LLM tem o seu próprio viés, que depende da sua arquitetura, processo de treino e datasets utilizados.

Mais especificamente no domínio clínico, surgem outras fontes de viés, também influenciadas pelo contexto e pelo utilizador, como se exemplifica na Figura 30 [93].

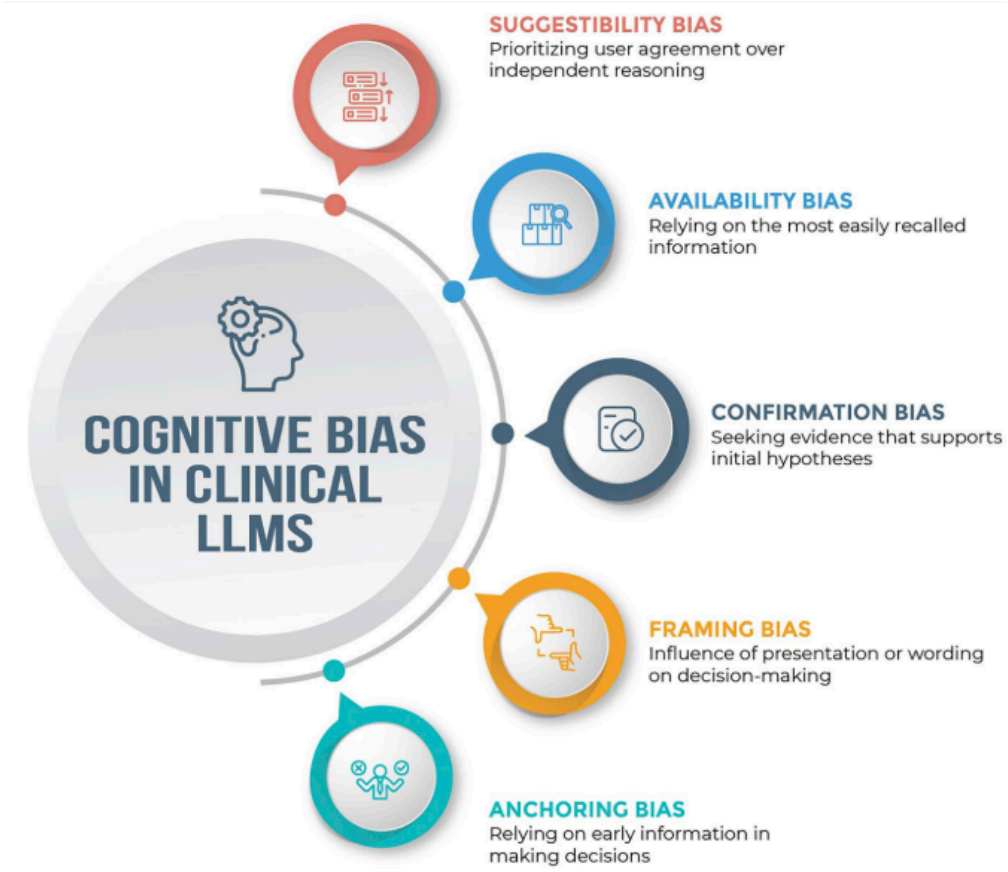


Figura 30 – LLMs: tipos de viés na prática clínica [93]

De todos estes, o *framing bias* e o *anchoring bias* são relevantes porque dependem apenas do utilizador, levantando a questão da formação das equipas clínicas para a correta interação com esta tecnologia [93].

Se a integração de LLMs na prática clínica acabar por ser generalizada, é lícito exigir formação para uma correta utilização.

5.5.3 Interpretabilidade

Os LLM funcionam como "caixas-negras"; ou seja, apesar de tecnicamente complexos e baseados em regras matemáticas, é difícil estabelecer uma visão retrospetiva que justifique determinado Output. A Interpretabilidade pode ser definida como a extração de conhecimento relevante do LLM sobre relações que estejam contidas nos dados ou tenham sido aprendidas pelo modelo [94].
Para uma análise deste problema, é necessário estudar a mecanística do modelo e obviamente o processo de treino [94], [95]. O decifrar deste problema pode passar também por compreender como o modelo atribuiu os pesos aos diferentes tokens, simplesmente "questionando" o próprio modelo [96].

5.5.4 Contexto e Memória

Cada modelo tem um limite para a quantidade de tokens que pode processar para gerar o(s) próximo(s) tokens – janela de contexto. Se a prompt inserida for demasiado longa, pode ser truncada e dados importantes do contexto não serem considerados. Este problema pode ser muito relevante quando a tarefa envolve o resumo de um artigo, ao ser pedido ao LLM para processar textos muito mais extensos do que aqueles que podem ter sido utilizados no treino.

Um outro problema envolve o escalar de recursos computacionais necessários. Alguns trabalhos recentes resumem as diferentes estratégias para ampliar a janela de contexto em LLMs: atenção eficiente, compressão de memória (resumir estados antigos), positional encoding avançado, memória externa e arquiteturas híbridas (Transformer + RNN) [97].

5.5.5 Segurança

Os LLM podem ser afetados por prompt adversarial - técnica utilizada para testar ou explorar vulnerabilidades, criando prompts concebidas para confundir, induzir em erro ou manipular os resultados do modelo.

Esta abordagem envolve a criação de prompts que desafiam a compreensão, o processo de tomada de decisão ou os limites éticos do modelo. Um deles é o efeito Waluigi – após ter sido treinado numa determinada direção ou comportamento, é mais simples induzir exatamente o oposto. Os principais tipos de prompt adversarial são: *jailbraking* (é feito um pedido para ignorar restrições de segurança); *prompt injection* (no texto da prompt são colocadas pequenas instruções que podem levar a comportamentos desviantes), etc.

5.5.6 Privacidade e Legalidade

Atendendo ao crescente uso de LLMs na Medicina, a manipulação de dados dos registos clínicos é altamente sensível e levanta desafios de privacidade. Regulamentações nacionais como o RGPD devem proporcionar estruturas de proteção, mas são necessárias estratégias específicas para mitigar os riscos na IA generativa.

Os riscos podem ser reduzidos usando estratégias como LLMs implantados localmente, sem conexão exterior, para preservar quer a privacidade, quer a segurança. Se de todo necessária a conexão exterior com LLMs state of the art e RAG, então devem ser observadas todas as regras para a anonimização dos dados [67], [72], [98].

Em última análise, o utente deve ser sempre o decisor sobre a utilização dos seus dados neste tipo de soluções [99].

5.5.7 Outras limitações

Já foi referido o problema do Esquecimento Catastrófico e do Overfitting, bem como toda a problemática da incerteza. Resta o problema material do custo computacional e hardware necessários para executar modelos mais complexos offline, e o custo inerente a toda a equipa dedicada a este (novo) sistema, bem como o custo da adaptação das diversas unidades clínicas.

Objetivamente, a integração da IA permitiria ir mais além do que apenas trabalhar com apoio de LLMs dedicados.

5.6 Aplicabilidade e Exequibilidade

Um LLM necessita de ter a capacidade de processar convenientemente a *prompt* inserida. Um dos fatores mais limitativos é o número de parâmetros, que atualmente se mede em biliões de parâmetros. Mas um maior número de parâmetros necessita de uma maior capacidade de memória em tempo real. Com esta progressão, rapidamente se ultrapassa o hardware mais básico e o custo real pode aumentar exponencialmente.

Existem diversas soluções *cloud-based*, com custos variados, passíveis de serem utilizadas em contexto de cuidados de saúde. Mas todas necessitam que os dados, apesar de anonimizados, sejam enviados para um outro sistema fora da rede local da unidade de saúde e, possivelmente, fora do país ou espaço europeu. Todas estas passagens aumentam exponencialmente os riscos de segurança.

6 SETUP EXPERIMENTAL DO PROJETO

Resumidamente, pretende-se utilizar um modelo LLM para processar textos clínicos de uma UCIN, com o objetivo de extrair dados objetivos, compará-los com referências, identificar diferenças e produzir uma solução para ajuste, sob a forma de uma proposta de ajuste nutricional.

Para que esse processo possa ser implementado é necessário um dataset, quer com dados reais, quer com dados sintéticos. Na ausência de um dataset real, optou-se por um sintético. Em virtude de não estar disponível um dataset sintético pré-feito, a única opção foi criar um dataset sintético *de novo* (abordagem detalhada: ver 6.3).

Foi também seguida uma abordagem sistematizada e racional para a análise e caracterização do processo deste Projeto, com base na Metodologia CRISP-DM [100], com uma adaptação específica pensada para dados de saúde – CRISP-MED-DM [101].

6.1 Dados – perspetiva CRISP-DM

Atendendo à complexidade dos dados, utilizou-se uma adaptação da metodologia CRISP-DM (Cross Industry Standard Process for Data Mining) [100] – esquematizado na Figura 31. Esta adaptação, proposta por Niaksu para o contexto médico, denomina-se CRISP-MED-DM [101].

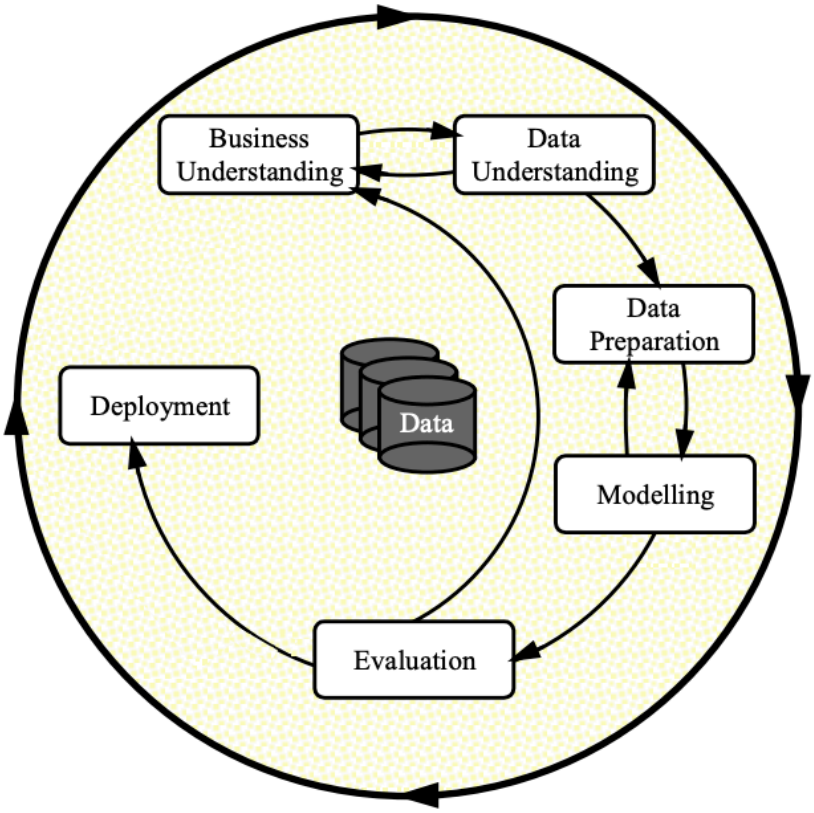


Figura 31 - Esquema clássico da metodologia CRISP-DM [100].

A adaptação da metodologia CRISP-DM para o contexto médico através de uma nova abordagem (CRISP-MED-DM) é justificada por diversos desafios específicos [101]:

- Grande variedade de representações e formato dos dados (bases de dados múltiplas, imagem, texto, vídeo, etc.) que exigem pré-processamento complexo e programação.
- Heterogeneidade dos dados – em medicina, o mesmo conceito / entidade pode ter nomes diferentes e identificadores múltiplos consoante o classificador utilizado. Pode ser necessário a utilização de uma ontologia comum. Por outro lado, a interoperabilidade dos diversos sistemas nas unidades clínicas não

significa que todos eles suportem ou usem uma norma comum como HL7[5] ou DICOM[6], sendo frequente a sua conversão posterior.

- Privacidade dos dados e RGPD[7]: a legislação não permite a utilização de dados clínicos sem a permissão expressa do utente ou seu representante legal. Para que possam utilizados, devem ser posteriormente anonimizados.
- Qualidade e integridade dos dados clínicos: frequentemente os registos apresentam dados incompletos e ainda os erros inerentes quer ao registo por pessoal de saúde, quer à própria medição pelo aparelho.

Globalmente, podemos identificar algumas diferenças:

Fase 1 – Business Understanding

Business understanding passa a Clinical Understanding para separar a aplicação no contexto clínico da perspetiva "business" mais ligada à economia da Saúde.

Os Objetivos foram divididos em objetivos clínicos e objetivos de gestão em economia da Saúde.

Foi adicionado um sub-item na caracterização (Assess Situation) especificamente para a Privacidade dos Dados e Aspetos Legais e outro para avaliar a integridade dos dados e das fontes. Estas adaptações estão referenciadas com mais pormenor na Figura 32.

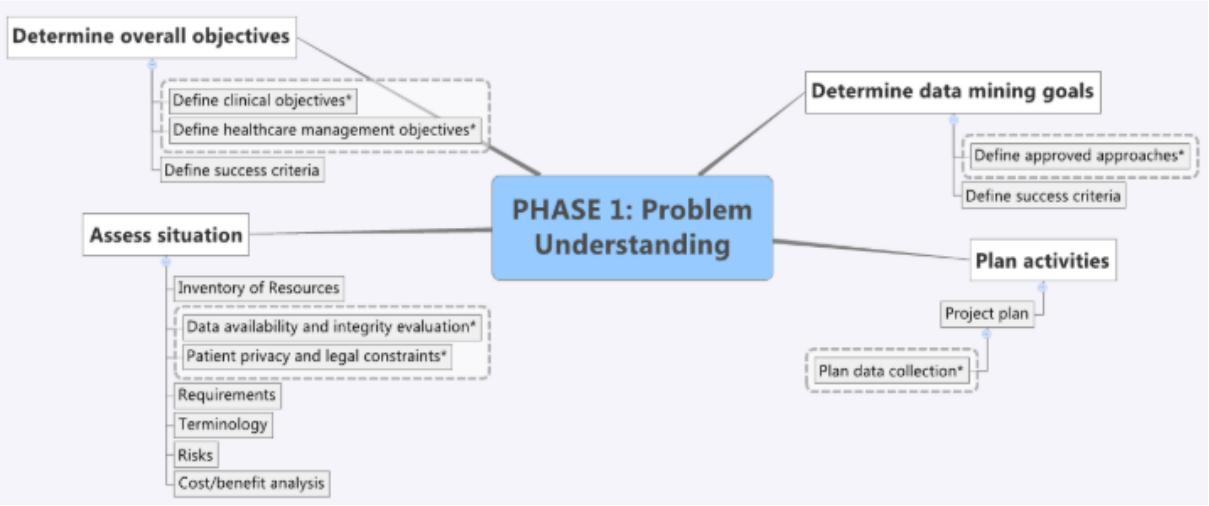


Figura 32 – Fase 1 com as adaptações propostas [100].

[5] HL7 – Health Level 7
[6] DICOM – Digital Imaging and Communications in Medicine
[7] RGPD – Regulamento Geral de Proteção de Dados

Fase 2 – Data Understanding

Nesta fase foi adicionada uma atividade – Preparar para a colheita de dados, que engloba a escolha das estratégias a seguir para o pré-processamento dos dados em formato diverso.

É também referido que devem selecionadas as ontologias e nomenclaturas a utilizar. Se clinicamente relevante, devem ser integrados os protocolos clínicos em utilização nas instituições pois influenciam o modo como os dados são registados. Estas adaptações estão referenciadas com mais pormenor na Figura 33.

Figura 33 – Fase 2 com as adaptações propostas [100].

Fase 3 – Data Preparation

Nesta fase existem também algumas diferenças, sobretudo relacionadas com a atividade original "Select Data", pois a maioria dos dados clínicos não estão em tabela de dupla entrada ou equivalente.

Foi também introduzida uma nova atividade, o "Prepare Data": implica a definição das interfaces para os variados sistemas das instituições de saúde; preparar o mapeamento das terminologias médicas; integrar a colheita de dados nos protocolos das instituições. Estas adaptações estão referenciadas com mais pormenor na Figura 34.

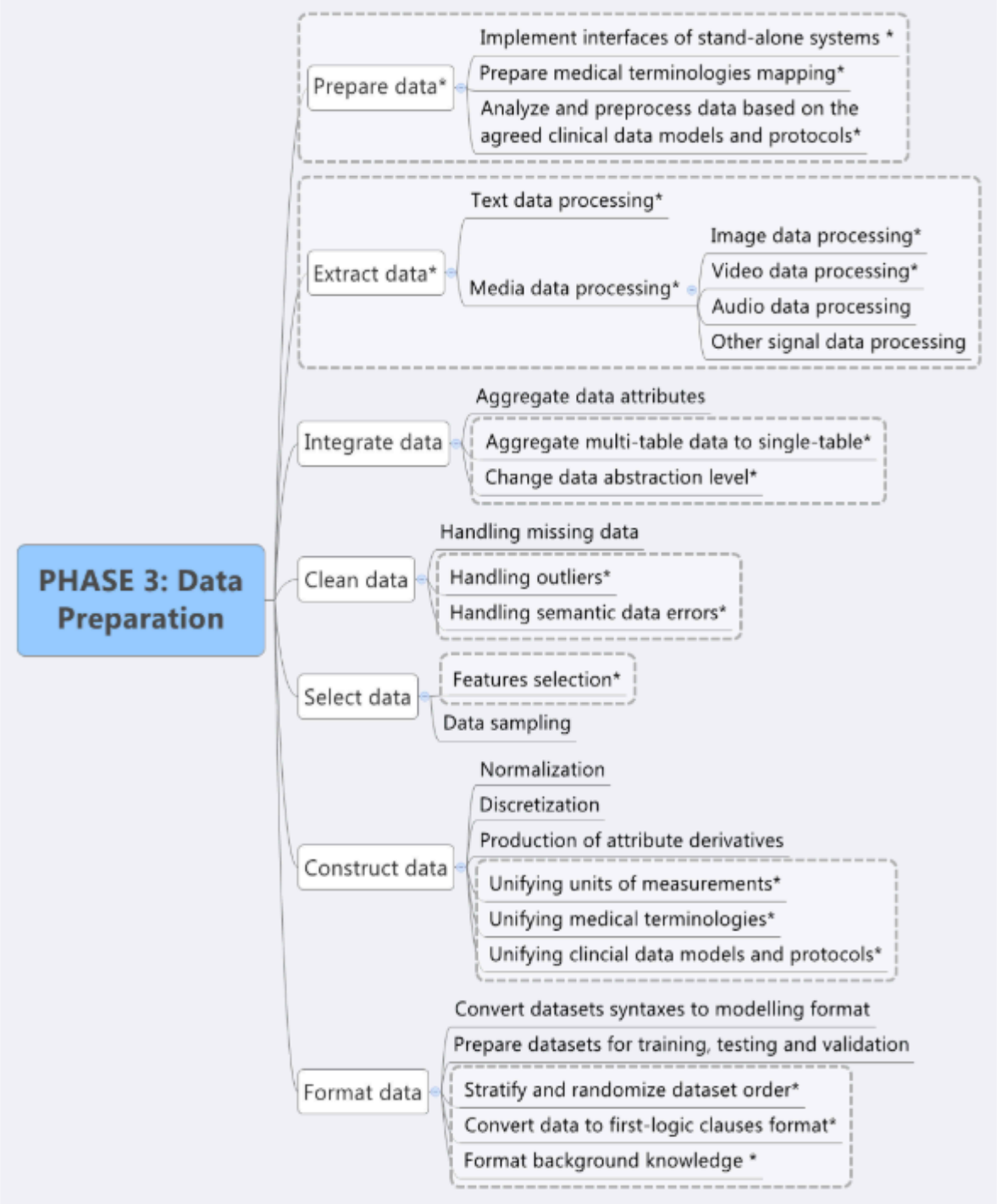


Figura 34 – Fase 3 com as adaptações propostas [100].

Fase 4 – Modelagem

Nesta fase não foram introduzidas atividades novas. Existe apenas uma nota sobre a possibilidade de exportação de dados de volta para o Processo Clínico Eletrónico, como se demonstra na Figura 35.

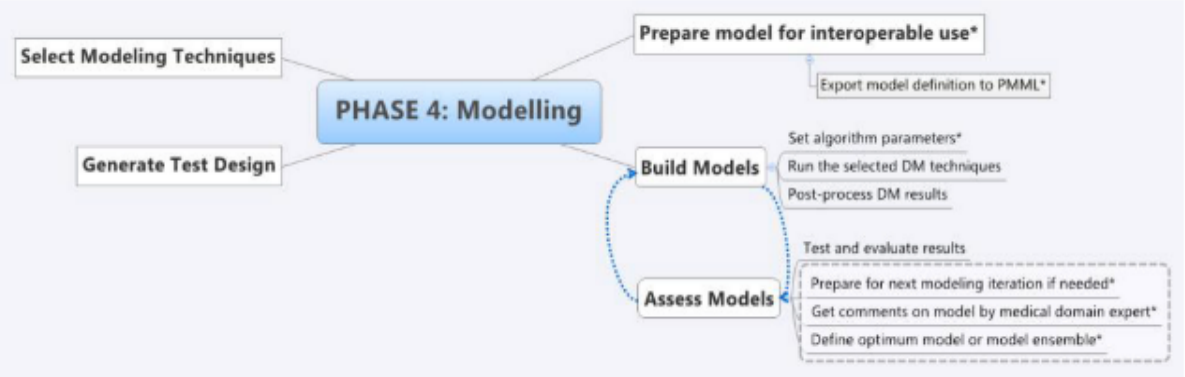


Figura 35 – Fase 4 de acordo com CRISP-MED-DM com as adaptações propostas [100].

Fase 5 – Avaliação

De acordo com a abordagem proposta segundo o CRISP-MED-DM, apenas foi introduzida a necessidade de comparar os resultados com outros estudos alternativos e com o gold standard (a referência reconhecida), como se pode verificar na Figura 36.

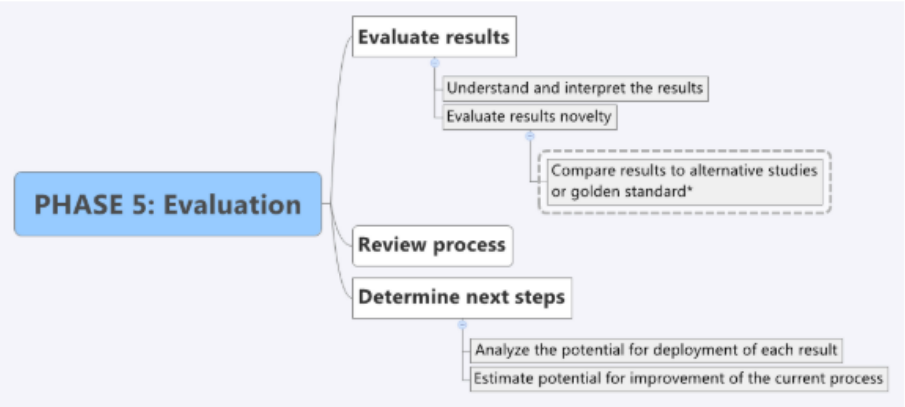


Figura 36 – Fase 5 com as adaptações propostas [100].

Fase 6 – Implementação

Nesta fase, o CRISP-MED-DM não introduz nenhuma adaptação ou novas atividades.

Após esta breve introdução à metodologia CRISP-MED-DM e evidenciadas as diferenças mais importantes, apresentamos o Projeto de acordo com este referencial.

6.2 Apresentação do Projeto da Tese

6.2.1 Clinical Understanding (Business Understanding)

Objetivos Clínicos (principais): utilização de IA; caracterizar crescimento e adequação nutricional em prematuros; propor ajustes nutricionais individualizados.

Objetivos de Gestão em Saúde (secundários): melhorar a qualidade dos cuidados; diminuir o desperdício de suplementos alimentares; diminuir o tempo médio de internamento.

Critérios de Sucesso: a) extração correta de dados relevantes dos diários clínicos; b) avaliação correta dos desvios em relação aos standards nutricionais; c) elaboração de uma proposta clinicamente plausível e adequada para correção desses desvios

População: RN internados na UCI neonatal.

Disponibilidade e Integridade dos dados: os dados estão disponíveis nos textos dos diários clínicos dos RN internados na UCIN. Como se trata de dados submetidos a processamento humano, poderão conter erros, valores ausentes ou valores absurdos. Alternativamente, poderão ser utilizados textos clínicos sintéticos, em tudo mimetizando os reais.

Privacidade e Restrições Legais: na utilização dos dados dos textos clínicos reais serão cumpridos todos os requisitos legais previstos no RGPD, bem como os pareceres das Comissões de Ética e de Proteção de Dados.

Foi efetuado pedido à Comissão de Ética da ULS, que remeteu para parecer do Encarregado de Proteção de Dados (DPO) da ULS.

No caso do dataset sintético, os pareceres acima referidos estão dispensados.

Requisitos, Custos e Constrangimentos: todo o projeto foi elaborado na perspetiva de uma dissertação de Tese de Mestrado, encontrando-se os requisitos e os custos integrados na mesma. Como constrangimentos, deve ser referida a dificuldade em ter acesso a dados reais e a opção pelo dataset sintético. Deve ainda ser referido a natureza experimental e não testada da real capacidade de um LLM offline conseguir cumprir com os objetivos propostos.

Terminologia (dados): foram utilizadas as definições clínicas universalmente aceites, e mapeadas para padrão internacional (SNOMED-CT para entidades clínicas e LOINC para entidades laboratoriais).

Variáveis: na Tabela 1 estão indicadas todas as variáveis a extrair das fontes de dados, com respetivos códigos LOINC e mapeamento para SNOMED-CT.

Tabela 1 - Lista de Variáveis, com mapeamento LOINC e SNOMED-CT para definir terminologia e semântica.

Sexo	Masculino Feminino	SNOMED-CT 248153007 SNOMED-CT 248153002
Cronológicas	Data de Nascimento Idade gestacional (semanas+dias) Dia de vida (D) Idade corrigida / IPM (semanas+dias)	LOINC: 21112-8 = Birth date SNOMED CT: 57036006 = Gestational age LOINC: 11884-4 = Gestational age at birth SNOMED CT: 427176004 = Day of life (observable entity) SNOMED CT: 715450001 = Postmenstrual age (observable entity)
Somatométricas	Peso (g) Comprimento (cm) Perímetro Cefálico (cm) Variação de Peso Z-Scores	SNOMED CT: 27113001 = Body weight LOINC: 3141-9 = Body weight Measured SNOMED CT: 8302-2 = Body height LOINC: 8302-2 = Body height, Stature SNOMED CT: 363812007 = Head circumference (observable entity) LOINC: 8287-5 = Head Occipital-frontal circumference Variável Calculada (Peso; valor absoluto e percentual) LOINC 8289-1 = Body weight-for-age Z-score LOINC 8288-3 = Body length-for-age Z-score LOINC 8287-5 = Head circumference-for-age Z-score
Nutricionais	Aporte hídrico total (ml/kg/dia) Calorias totais (kcal/kg/dia) Proteínas (g/kg/dia) Lípidos (g/kg/dia) Hidrat. carbono (g/kg/dia) Ferro (mg/kg/dia)	SNOMED CT: 225390008 = Fluid intake (observable entity) LOINC: "Fluid intake total per day" (LOINC: 3138-5) SNOMED CT: 443883004 = Energy intake (observable entity) SNOMED CT: 226356009 = Protein intake (observable entity) SNOMED CT: 226358005 = Fat intake (observable entity) SNOMED CT: 226357000 = Carbohydrate intake (observable entity) SNOMED CT: 102600008 = Iron intake (observable entity)

Riscos Expectáveis: alguns dados poderão não estar legíveis (notação incompreensível), ausentes ou com erros apontando para valores não fisiológicos (exemplo: peso = 30 000g); o modelo pode não conseguir extrair as variáveis pretendidas na totalidade; o dataset sintético pode não ser o mais fidedigno possível da realidade. Há que considerar ainda todos os riscos inerentes ao próprio LLM (ver Capítulo 5).

6.2.2 Data Understanding

O modelo vai analisar diários clínicos de uma unidade de cuidados intensivos neonatais (UCIN). Estes diários podem ser classificados como texto semi-estruturado: o texto médico obedece a uma determinada organização, um padrão relativamente estável.

Dicionário, Interoperabilidade Semântica: já referido na lista de variáveis e terminologia (ver Tabela 1 - Lista de Variáveis, com mapeamento LOINC e SNOMED-CT para definir terminologia e semântica.); estão assim definidas as nomenclaturas, classificadores e ontologias usadas.

Fontes de dados: diários clínicos; amostragem com periodicidade diária (SClínico© via API se disponível) e Dataset Sintético com textos similares.

Conjunto de Dados

Para que fosse facultado o acesso a dados reais, constantes de diários clínicos de RN internados numa UCIN, foi efetuado, atempadamente, pedido à Comissão de Ética que remeteu posteriormente para o Encarregado de Proteção de Dados da ULS, dos quais aguardamos, ainda, parecer (favorável ou desfavorável) à data de entrega desta Tese.

Para este projeto foi necessário utilizar um dataset sintético. Ora tal dataset não está, facilmente, disponível; quer pela sua especificidade e complexidade, quer pela barreira linguística (pretende-se que todo o Projeto opere em língua portuguesa). Apesar de vários esforços, não se conseguiu encontrar tal dataset em língua portuguesa.

Decidiu-se assim optar pela produção completa e autónoma de um dataset sintético apropriado, replicando o contexto clínico da UCIN e em português.

Todos os pormenores sobre a criação do mesmo estão descritos em 6.3 Dataset Sintético. Neste momento descrevemos apenas os aspetos gerais.

No final, obtivemos um dataset com 142 textos de exemplo de um diário clínico de RN em UCIN, semelhantes ao ilustrado na Figura 37.

Este é um texto semi-estruturado: obedece a uma determinada organização, um padrão relativamente estável. É este o padrão em utilização na UCIN. De uma maneira geral, todas as variáveis estão presentes no texto. A sua dimensão (número de palavras) é igualmente representativa dos textos reais.

XXNOMEXX, D30 de vida
IG 26S+2d | IPM 30s+4d
PN 905g | PA 1020g (-15g/24h)

LISTA DE PROBLEMAS:
Prematuridade | EBPN
SDR - NIPPV
A/B
Anemia desde o nascimento (TGV 7/3)
Má progressão ponderal
Hiponatrémia

Aportes: 168mL/Kg/dia (calorias 142kcal/kg/dia, proteinas 4,5g/Kg/dia, lipidos 6.4g/Kg/dia, PER 3.2g/100Kcal, Ferro 5mg/kg) | Na 2.7mEq/Kg/dia
Dispositivos: VNI, SOG

RESPIRATÓRIO: Em NIPPV desde D1, FiO2 21%. Instável com várias A/B, necessidade de O2 e estimulo para recuperar (rastreios septicos neg), mas noção de maior estabilidade após TGV ontem. Sob cafeína oral. GSV em D27 (05/03): pH 7.25, pCO2 51,6 mmHg, HCO3 22,7 mmol/L, EB -45 mmol/L, AG 16,6 mmol/L.

CARDIOVASCULAR: Hemodinamicamente estável. Último lactato D27 (05/03): 5 mmol/L. Ecocardio funcional D29 (7/3): boa função ventriular, FE 45%, Ae/Ao 1,2. FOP. Sem PCA aparente.

RENAL: Diurese regular, não contabilizada. Última função de D27 (05/03): creatinina 0,5 mg/dL e azoto ureico 11mg/dL.

DIGESTIVO: AET desde D8 (14/2), com tolerância. Suplemento proteico desde D20 e FMS desde D21. Dejeções espontâneas e regulares. ABD normal.

METABÓLICO: Normoglicemia desde D20 (26/2). Última GSV em D27: Na 139 mmol/L, K 5 mmol/L, Cl 104 mmol/L, Ca(i) 5.4mg/dL. Sob suplementação com Na+ oral: 2,7meq/Kg/dia. Rastreio DMO-PT em D27 normal.

HEMATOLÓGICO: Anemia desde D0, último hemograma em D27: Hb 8,4g/dL, Htc 24%, Leuc 14100, Plaq 1.019.000/uL, Ret 240.000/uL - decididio em equipa fazer TGV que fez a D29 (7/3), sem intercorrências. Sob ferro oral desde D14 (20/02), atualmente 5mg/Kg/dia.

INFECIOSO: Sem antibioterapia. Último rastreio sético em D27 negativo. Rastreios ESBL seriados negativos, último a 03/03.

NEUROLÓGICO: Tónus e postura adequados à IG. Última EcoTF D21(27/2): sem LMQ. Ligeira hiperecogenecidade do núcleo caudado bilateralmente (menos que o plexo coróide). Doppler da ACA c/ IR de 0.74.

OUTROS: DP nº 1 em D5 (11/02) - normal e nº2 em D14 (20/02). Somatometria D21: peso 875g, comp 34,7cm, PC 22,1cm.

PLANO:
- GSV.
- Somatometria amanhã.

Figura 37 - Exemplo genérico de texto a utilizar (NB – dados fictícios)

Exploração inicial [8]:

O dataset criado e utilizado neste projeto apresenta as características que se encontram resumidas na Tabela 2.

Tabela 2 – Estatística Descritiva do Dataset

	Validas	Missing	Média	Mediana	DP	Min	Q1	Q3	Max
Idade Gestacional (s)	141,00	1,00	27.94	27,00	2.31	24,00	27,00	28,00	36,00
Peso de Nascimento (g)	141,00	1,00	1140.38	1020,00	450.46	632,00	944,00	1150,00	3250,00
Comprimento (cm)	96,00	46,00	36.308	35.8	2.39	31.9	34.3	38.92	40.5
Perímetro Cefálico (cm)	95,00	47,00	26.143	26.5	3.09	21,00	23.05	28.4	35.5
Aporte hídrico total (ml/kg/dia)	133,00	9,00	154.508	158,00	13.17	100,00	150,00	163,00	183,00
Calorias Totais (kcal/kg/dia)	129,00	13,00	129.16	133,00	14.22	84,00	125,00	138,00	165,00
Proteínas (g/kg/dia)	129,00	13,00	3.82	4,00	0.51	1.4	3.7	4.1	4.4
Lípidos (g/kg/dia)	124,00	18,00	5.94	6.1	0.70	4,00	5.6	6.4	8.9
PER (g/100 kcal)	119,00	23,00	3.02	3,00	0.26	2.5	2.9	3.1	4,00
Ferro (mg/kg/dia)	123,00	19,00	4.64	5.1	1.33	0,00	4,00	5.45	6.5

Nota: Total de Casos (textos de diários clínicos) = 142 (56 masculinos; 86 Femininos)

Este dataset engloba RN de Extrema Prematuridade e também alguns RN de Prematuridade Tardia, como tentativa de replicação de um internamento normal numa UCIN (ver Figura 38) e as principais características desta população (ver Figura 39 e Figura 40).

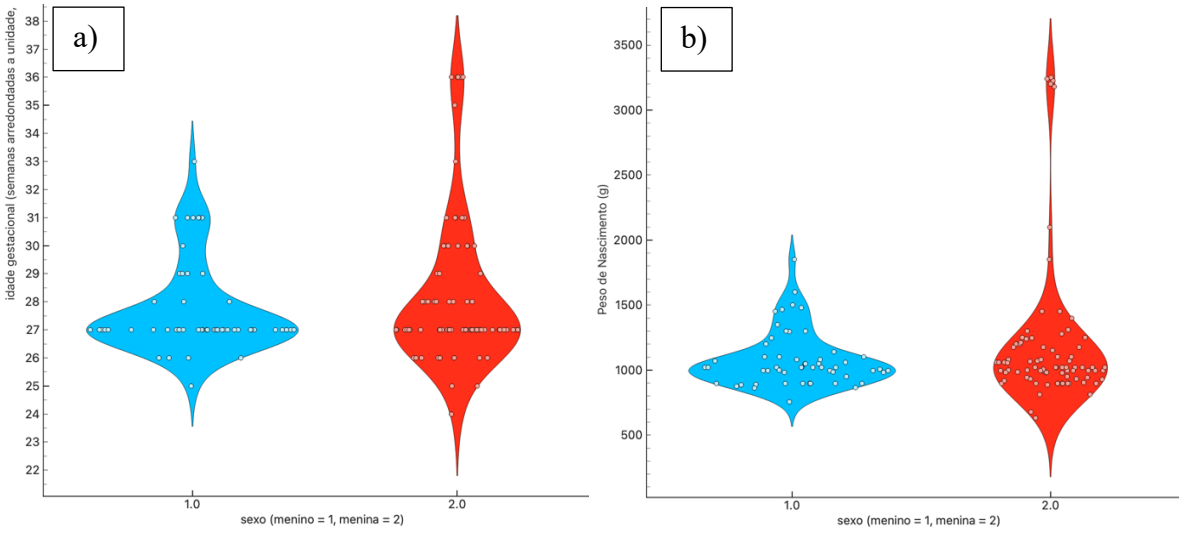


Figura 38 – Violin Plot: a) distribuição de idade gestacional por sexo (note-se a distribuição com maior frequência em torno das 27s e depois novamente, mas em menor número em torno das 32s e superiores); b) raciocínio idêntico, mas com o peso de nascimento.

[8] Na exploração geral do dataset foi utilizado o Orange Data Mining. (2025). Orange (versão 3.3.8). https://orangedatamining.com (último acesso em 4 de outubro de 2025)

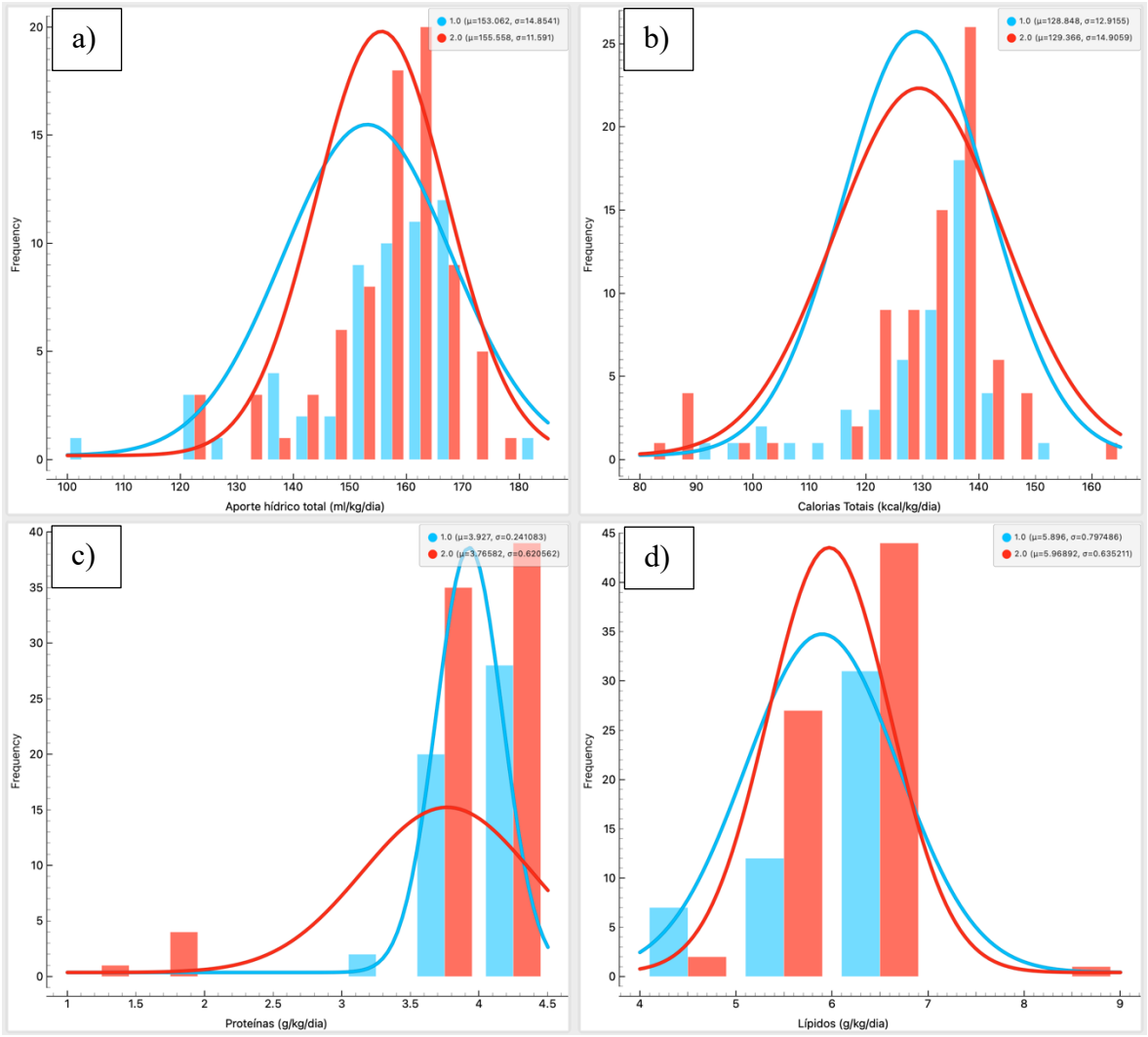


Figura 39 – Distribuição por sexo de a) Aporte Hídrico Total, b) Calorias Totais, c) Proteínas e d) Lípidos.

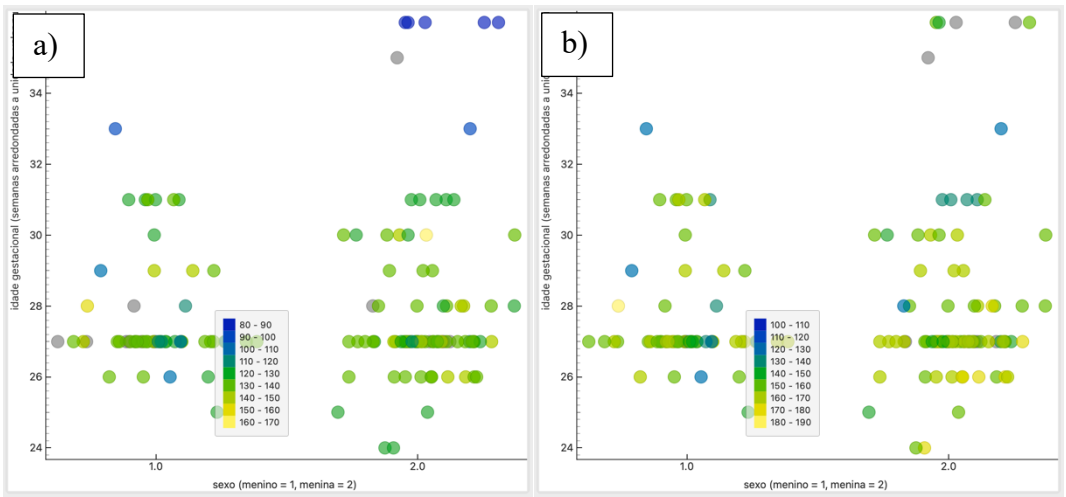


Figura 40 – Distribuição por sexo e idade gestacional: a) Aporte Hídrico Total (ml/kg/dia) e b) Aporte Calórico Total (kcal/kg/dia)

Qualidade dos dados

Após análise das tabelas e figuas anteriores, podemos constatar que o dataset possui uma distribuição compatível com o teor dos internamentos em UCIN, e que a somatometria (peso, comprimento e perímetro cefálico) acompanham a idade gestacional; bem como os aportes (hídrico total, calorias, proteínas e lípidos) estão de acordo com o esperado.

Globalmente, não são assinaláveis desvios em relação a valores fisiológicos esperados.

Alguns valores em falta (missing) foram intencionalmente criados para melhor traduzir a possibilidade de registos incompletos (tal como um dataset real). Não existem outliers assinaláveis, provavelmente devido ao processo de geração do dataset sintético.

6.2.3 Data Preparation

Para esta fase, devem ser considerados os seguintes itens:

- Necessidade de definição de intervalos fisiológicos
- **Normalização de unidades** (conformidade LOINC ©)
    - ml/kg/dia, g/kg/dia, kcal/kg/dia.
- **Cálculos / variáveis derivadas**
    - Dia de Vida (se não estiver registado) = data de observação – data de nascimento
    - Idade corrigida / IPM (se não estiver registado)
    - Conversão de semanas + dias em número absoluto
    - Variação de peso em valor absoluto e percentagem em relação ao peso de nascimento
    - Z-scores (tabelas Intergrowth 21st)
- **Exportação dos dados**
    - Tabela com as variáveis iniciais e derivadas

6.2.4 Modelagem

Nesta fase pretende-se idealizar a melhor opção para extração e manipulação dos dados e alguns cálculos acessórios.
Implica a preparação do modelo com uma prompt ajustada e eficaz.
Neste ponto, apenas se poderá propor o arredondamento dos cálculos para as semanas de idade corrigida completas para alguns parâmetros.

- Vantagens: facilita os cálculos pelo LLM e a sua comparação com os standards nutricionais e tabelas de referência (Anexos 3 e 4); permite raciocinar com base em Z-Scores.
- Desvantagens: perda de granularidade dos dados (não significativa).

6.2.5 Avaliação

Teste do modelo:
Nesta fase estaremos a testar a capacidade do modelo em extrair dados, verificar desvios e propor ajustes aos aportes nutricionais.

6.2.6 Clinical Deployment

Objetivo final do projeto, depois do teste. Atendendo aos constrangimentos no acesso aos dados reais, esta etapa será desenvolvida em trabalho futuro.

6.3 Dataset Sintético

6.3.1 Metodologia de Geração Sintética dos Textos

A construção do dataset baseou-se na geração controlada de texto clínico com recurso a modelos de linguagem de grande escala (LLMs), utilizados de forma comparativa para avaliar consistência interna, coerência médica e adequação terminológica.
Foram testados cinco modelos com diferentes arquiteturas e modos de execução: GPT-5.0 (OpenAI©), Gemini Advanced 1.0 (Google©), DeepSeek V2©, Grok-2.0 (xAI©) e Manus AI 1.0© [9].

Diagrama do Processo:

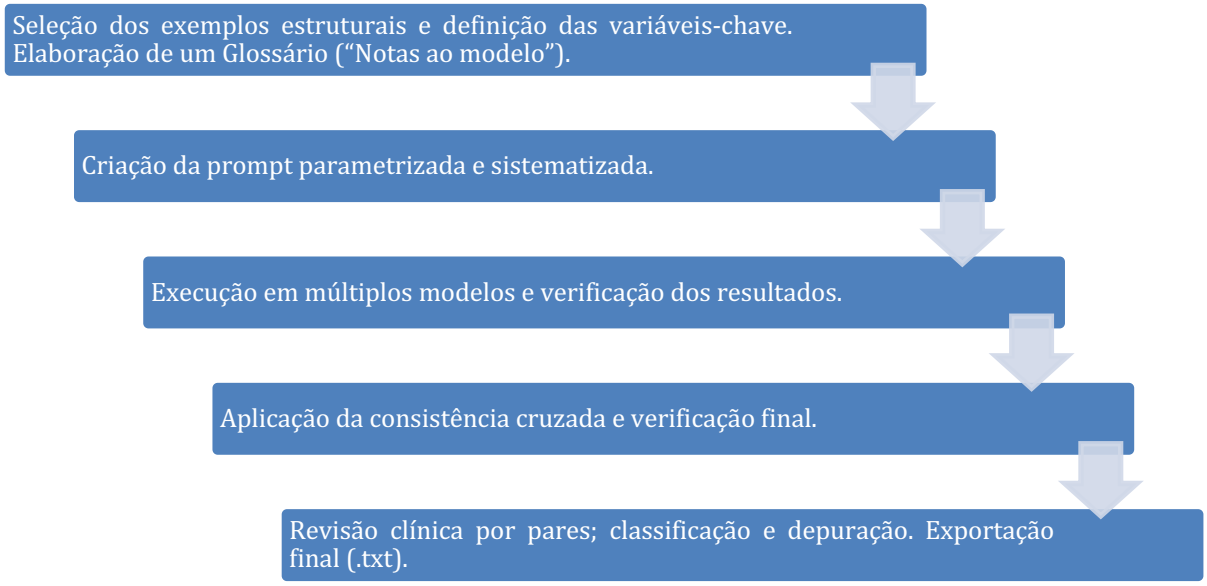


Figura 41 – Esquema global do processo

[9] Estes modelos funcionam exclusivamente online, englobam as versões de acesso livre e, constituem as versões atualizadas e disponíveis à data do processo de geração do dataset sintético, ou seja, durante o período de junho a setembro de 2025. Neste período podemos adiantar alguns dados sobre os modelos, não completamente confirmados: GPT-5.0 – 1,5-3,0 Triliões de Parâmetros, arquitetura MoE complexa; Gemini Advanced 1.0 – 1-2 Triliões de Parâmetros, arquitetura MoE multimodal; DeepSeek V2 – 250-300 Milhares de Milhões de Parâmetros, MoE eficiente; Grok-2.0 – 300-500 Milhares de Milhões de Parâmetros, MoE parcial (o mais denso de todos); Manus AI 1.0 70-100 Milhares de Milhões de Parâmetros (ligeiramente diferente no sentido em que se trata de um Agente utilizando um LLM).

6.3.2 Abordagem Geral

O processo iniciou-se a partir de exemplos realistas e plausíveis de diários clínicos de unidade de cuidados intensivos neonatais, criados pelo autor e por um conjunto de pares (três neonatologistas da ULS).

Foi criado um conjunto inicial de 25 exemplos, originais, como o exemplo na Figura 42.

Cada exemplo foi depois integrado num prompt parametrizado, com instruções explícitas relativas a idade gestacional, peso de nascimento, idade pós-menstrual e aportes nutricionais esperados.

A geração ocorreu em ciclos iterativos, avaliando-se coerência clínica, progressão temporal e uniformidade terminológica. Cada nota gerada foi revista e, quando necessário, ajustada para assegurar plausibilidade fisiológica.

XXNOMEXX, D30 de vida
IG 26S+2d | IPM 30s+4d
PN 905g | PA 1020g (-15g/24h)

LISTA DE PROBLEMAS:
Prematuridade | EBPN
SDR - NIPPV
A/B
Anemia desde o nascimento (TGV 7/3)
Má progressão ponderal
Hiponatrémia

Aportes: 168mL/Kg/dia (calorias 142kcal/kg/dia, proteinas 4,5g/Kg/dia, lipidos 6.4g/Kg/dia, PER 3.2g/100Kcal, Ferro 5mg/kg) | Na 2.7mEq/Kg/dia
Dispositivos: VNI, SOG

RESPIRATÓRIO: Em NIPPV desde D1, FiO2 21%. Instável com várias A/B, necessidade de O2 e estimulo para recuperar (rastreios septicos neg), mas noção de maior estabilidade após TGV ontem. Sob cafeína oral. GSV em D27 (05/03): pH 7.25, pCO2 51,6 mmHg, HCO3 22,7 mmol/L, EB -45 mmol/L, AG 16,6 mmol/L.

CARDIOVASCULAR: Hemodinamicamente estável. Último lactato D27 (05/03): 5 mmol/L. Ecocardio funcional D29 (7/3): boa função ventriular, FE 45%, Ae/Ao 1,2. FOP. Sem PCA aparente.

RENAL: Diurese regular, não contabilizada. Última função de D27 (05/03): creatinina 0,5 mg/dL e azoto ureico 11mg/dL.

DIGESTIVO: AET desde D8 (14/2), com tolerância. Suplemento proteico desde D20 e FMS desde D21. Dejeções espontâneas e regulares. ABD normal.

METABÓLICO: Normoglicemia desde D20 (26/2). Última GSV em D27: Na 139 mmol/L, K 5 mmol/L, Cl 104 mmol/L, Ca(i) 5.4mg/dL. Sob suplementação com Na+ oral: 2,7meq/Kg/dia. Rastreio DMO-PT em D27 normal.

HEMATOLÓGICO: Anemia desde D0, último hemograma em D27: Hb 8,4g/dL, Htc 24%, Leuc 14100, Plaq 1.019.000/uL, Ret 240.000/uL - decididio em equipa fazer TGV que fez a D29 (7/3), sem intercorrências. Sob ferro oral desde D14 (20/02), atualmente 5mg/Kg/dia.

INFECIOSO: Sem antibioterapia. Último rastreio sético em D27 negativo. Rastreios ESBL seriados negativos, último a 03/03.

NEUROLÓGICO: Tónus e postura adequados à IG. Última EcoTF D21(27/2): sem LMQ. Ligeira hiperecogenecidade do núcleo caudado bilateralmente (menos que o plexo coróide). Doppler da ACA c/ IR de 0.74.

OUTROS: DP nº 1 em D5 (11/02) - normal e nº2 em D14 (20/02). Somatometria D21: peso 875g, comp 34,7cm, PC 22,1cm.

PLANO:
- GSV.
- Somatometria amanhã.

Figura 42- Exemplo genérico de texto a utilizar (dados fictícios)

6.3.3 Estratégia de *Prompting*

Foram testadas duas abordagens principais:

- Single-shot generation – o modelo produzia um diário completo a partir de um único prompt e conjunto de variáveis. Esta abordagem revelou-se a mais estável, garantindo coerência interna e evitando deriva semântica.
- Multiple-shot generation – sequência de exemplos (ex.: dia 1, dia 3, dia 5), usada para reforçar consistência longitudinal em modelos com maior capacidade de contexto.

Embora os fabricantes declarem janelas de contexto muito extensas (de 32 000 a mais de 1 000 000 tokens), na prática a maioria dos modelos apenas permitiu processar um diário clínico real de referência por execução, apesar de se tratar das versões online e *state of the art*.

GPT-5 e Gemini Advanced toleraram multi-shot prompting limitado (até dois) sem perda relevante de coerência, embora com ganho modesto.

DeepSeek V2, Grok-2 e Manus AI apresentaram degradação progressiva da consistência quando incluídos múltiplos exemplos, exigindo uma abordagem single-shot mais parcimoniosa.

Esta limitação prática resultou menos da dimensão teórica da janela de contexto e mais da forma como cada modelo gere atenção e ponderação de relevância em textos longos com instruções complexas. Nomeadamente, houve perda significativa de complexidade e extensão e até frases truncadas e simplificadas (ver Figura 43).

RN 44
D30 de vida
IG 27s+6d | IPM 32s+1d
PN 995g | PA 1380g (+30g/dia)

DIAGNÓSTICOS/PROBLEMAS:
PT 27s+6d | EBPN 995g
DBP (leve)
Anemia

Dispositivos: SNG | CNAF (2L/min)
Aportes programados: AHT 165 mL/Kg/dia | Cal 145 Kcal/Kg/dia; Prot T 4.4 g/Kg/dia; Lip T 6.3 g/Kg/dia; PER 3.1 g/100 kcal | Na 1.8 mEq/Kg/dia | Fe 4.8 mg/Kg/dia | K 2.1 mEq/Kg/dia

EVOLUÇÃO:
RESPIRATÓRIO: Estável em CNAF. FiO2 21%.
CARDIOVASCULAR: HD estável.
DIGESTIVO: AET total.
METABÓLICO: Estável.
HEMATOLÓGICO: Hb 9.5 g/dL. Sem sintomas de anemia.
NEUROLÓGICO: Boa vitalidade.

PLANO:
- Tentar desmame do CNAF para AR ambiente.
- Aumentar aporte de Ferro para 5.8 mg/kg/dia.
- Programar avaliação oftalmológica.

Figura 43 – Exemplo de texto clínico sintético, considerado não conforme pela falta de complexidade e simplificação absurda. Deve ser referido que o parágrafo com a linha sobre aportes foi respeitada e encontra-se na posição habitual. Podemos ainda encontrar alguns exemplos de pormenores atribuíveis a um modelo generalista / não específico de medicina, como a notação DBP leve (classificação inexata) ou ainda a omissão de itens (Renal e Infecioso, por exemplo). Texto gerado com o DeepSeek©.

Para comparação, segue um exemplo praticamente indistinguível de um diário clínico real – ver Figura 44.

```
RN 88
D34 de vida
IG 27s+1d || IPM 32s+1d
PN 895g || PA 1460g (+25g/24h)

PROBLEMAS/DIAGNÓSTICOS:
# PT 27s+1d | EBPN 895g
# Suspeita de descolamento da placenta
# SDR com necessidade transitória de O2 – em melhoria
# CIA OS pequena
# Anemia
# Hipoclorémia subclínica
# Trombocitose

Dispositivos: SOG | nCPAP (PEEP 6,5 cmH2O, FiO2 22–25%)

Aportes programados:
AHT 150 mL/Kg/dia | cal 130,5 Kcal/Kg/dia; prot T 4,1 g/Kg/dia; Lip T 6,3 g/Kg/dia; PER 3,1 g/100
kcal Na 3,1 mEq/Kg/dia | Fe 5,6 mg/Kg/dia | K 2,1 mEq/Kg/dia

EVOLUÇÃO
RESPIRATÓRIO: Em nCPAP com pequenas oscilações de FiO2 22–25%. Sem A/B nas últimas 48 h. EcoLung
com redução de linhas B. Sob cafeína.
CARDIOVASCULAR: Estável. Sopro II/VI. Eco: CIA OS 3 mm, sem repercussão. Lactato 1,4 mmol/L.
RENAL: DU adequada. Função renal normal.
DIGESTIVO: Enteral total D13. FMS e suplemento proteico ativos. Sem resíduos. Abdómen depressível.
METABÓLICO: Ionograma: Na 136 mmol/L, K 4,0 mmol/L, Cl 95 mmol/L. BrbT 2,7 mg/dL.
HEMATOLÓGICO: Hb 8,4 g/dL, Ht 25,3%. Plaquetas 630 000/µL. Reticulócitos 295 000/µL.
INFECIOSO: Sem antibioterapia. Vigilância ESBL negativa.
NEUROLÓGICO: EcoTF com vasculopatia lenticuloestriada dta residual. Sem convulsões. Tónus adequado.
OFTALMOLÓGICO: Sem ROP. Próxima avaliação programada.
OUTROS: Somatometria D34: peso 1460 g, comp 38,8 cm, PC 28,1 cm. DP nº3 D31.

PLANO:
- Manter plano nutricional e cafeína.
- Repetir somatometria em 48–72 h.
- Reavaliar necessidade de cloro oral se Cl < 96 persistir.
```

Figura 44 – Diário Clínico Sintético, considerado de boa qualidade. GPT 5©.

Cada prompt incluía instruções normalizadas, reproduzidas na Figura 45:

Prompt

Sou neonatologista e necessito de criar notas clínicas sintéticas a partir de exemplos que vou fornecer. Podes fazer isso?
Preciso de mais 9 exemplares sintéticos a partir de cada um dos exemplos. As notas sintéticas devem ser um pouco mais variadas mas realistas. Nota bem, podes variar, mas os pesos atuais, os pesos de nascimento e a idade gestacional devem ser realistas. Podes usar, para esclarecer algum termo desconhecido, as referencias no documento "notas ao modelo" mas não necessitas de referir ou citar o mesmo.
Necessito que sigas a estrutura / layout de notas clínicas fornecida, em todos os exemplos que criares deve existir informação em cada subitem do layout.
No item "aportes programados", devem aparecer identificados os parcelares de nutrição como neste exemplo:"AHT 151mL/Kg/dia | cal 128,8 Kcal/Kg/dia; prot T 4,0g/Kg/dia; Lip T 5,8g/Kg/dia; PER 3,1 g/100 kcal Na 3 mEq/Kg/dia | Fe 5,8 mg/Kg/dia | K 2,1meq/kg/dia". Necessito disso em todos eles.

Figura 45 – Prompt utilizada em todos os modelos. Tal como explicitado ao modelo, foi fornecido um glossário em arquivo .txt adicional ("notas ao modelo.txt"), para esclarecer dúvidas de siglas e acrónimos comuns nos diários clínicos. Foram também fornecidos os exemplos originais.

Abaixo, na Figura 46, encontra-se um excerto do glossário fornecido com o nome de "notas ao modelo.txt".

AHT - aporte hídrico total
EPC - cateter epicutaneo - cava (PICC)
MID / MIE / MSD / MSE - membro inferior / superior / direito / esquerdo
GSV - gasimetria venosa
DU - debito urinário
ECG - electrocardiograma
Brb - bilirrubina
SOG / SNG - sonda orogástrica / nasogástrica
VNI - ventilação não invasiva
EBPN / MBPN / BPN - extremo / muito baixo / baixo peso ao nascimento
DMOP / DMOPT - doença metabólica óssea da prematuridade
PA - peso atual
HC - hemocultura
SGB - streptococcus grupo B
DR - diagnóstico retrospectivo
DBP - displasia broncopulmonar
DP - diagnóstico precoce
WES - wide exome sequence
CVU - cateter venoso umbilical
TGV - transfusão de glóbulos vermelhos
Ev - endovenoso
Po - per os (via oral)
HD - hemodinamicamente
C10 - teste rapido de urina
LU Score - lung ultrasound score
PER - protein to energy ratio

Figura 46 – "Notas ao modelo.txt" – pequeno glossário fornecido ao modelo para esclarecer as siglas e acrónimos utilizados.

No final, a estratégia de single shot foi a mais eficaz, gerando resultados de melhor qualidade. Em todos os modelos, provavelmente pela dimensão da *context window*, ao serem carregados dois ou mais textos de exemplo, os diários clínicos gerados continham menor densidade de texto e menos pormenor.

Apesar de pedido inicialmente 9 exemplos, a tarefa foi incremental e foi terminada aos 6 (sobretudo pela perda de qualidade nalguns modelos).

Deve ser referido que os termos utilizados no exemplo e obtidos nos textos criados integram-se com SNOMED-CT© e LOINC ©.

Uma nota sobre o desempenho dos modelos utilizados para gerar este dataset encontra-se referido no próximo ponto - Capítulo 6.3.4.

Obtivemos assim uma amostra total "bruta" de 175 diários clínicos, dos quais 25 eram os originais criados pelo autor.

6.3.4 Comparação entre Modelos

A utilização de modelos online para criar os textos do dataset sintético obriga a contextualizar o desempenho dos mesmos a uma determinada janela temporal (setembro-outubro de 2025).

Os modelos apresentaram comportamentos distintos quanto à coerência clínica, estabilidade linguística e tolerância ao multiple-shot prompting. Apesar de podermos especular e recolher alguns dados dos próprios modelos, os pormenores do treino de cada um, os contextos utilizados e também o modo "interno" de funcionamento obviamente influenciaram os diferentes resultados obtidos – ver Tabela 3 – Comparação Qualitativa dos diferentes LLMs utilizados.

Tabela 3 – Comparação Qualitativa dos diferentes LLMs utilizados

	Facilidade em cumprir instruções sem correção	**Tarefa Completada**	**One-shot vs Multiple Shot**	**Consistência Clínica**	**Atenção a intervalos fisiológicos**	**Consistência Cruzada**	**Qualidade geral do texto**
GPT 5	+++	+++	2-3 ex	+++	+++	+++	+++
Gemini 2.5	+++	+++	2 ex	+++	+++	++	++
Grok	++	+++	1-2 ex	++	++	++	++
DeepSeek V2	++	++	1 ex	+-	++	+	---
Manus AI	+++	+++	1-2 ex	++	+++	++	++

Exploração resumida dos resultados:

- GPT-5© (OpenAI)— Maior consistência clínica e capacidade de manter coerência temporal, mesmo com prompts longos. Suportou dois a três exemplos reais de referência sem degradação notável. Cumpriu os intervalos fisiológicos de aportes e exigiu mínima intervenção manual.
- Gemini Advanced 1.0 © (Google) — Produziu texto fluente e linguagem natural, mas com tendência à simplificação terminológica neonatal. Aceitou multi-shot limitado; acima de três exemplos, observou-se redução da precisão numérica e da adesão às instruções.
- DeepSeek V2© (DeepSeek) — Apesar da boa fluência sintática em prompts simples, demonstrou baixa densidade na informação clínica. Em vários casos, os campos da sistematização clínica (respiratório, cardiovascular, digestivo, neurológico) foram preenchidos com apenas três a quatro palavras genéricas, sem descrição evolutiva nem parâmetros quantitativos. Essa limitação sugere menor capacidade de retenção contextual. O modelo exigiu prompts mais detalhados e ainda assim manteve desempenho inferior, acabando por os textos serem considerados de baixa qualidade (C) e eliminados.

- Grok-2© (xAI) — Favoreceu criatividade textual e variação estilística, porém com menor rigor técnico. A coerência interna deteriorava-se rapidamente quando o prompt incluiu mais de um diário.
- Manus AI© (Butterfly Effect) — Plataforma que disponibiliza acesso a modelos genéricos (Gemma, Mistral, LLaMA). O desempenho foi variável entre sessões, refletindo a diversidade de arquiteturas subjacentes. Apesar da boa fluência linguística, demonstrou limitação prática de contexto e apenas single-shot prompting estável. Requereu prompts mais detalhados para manter coerência clínica, dado não estar otimizado para terminologia médica. Não demonstrou, de modo objetivo, diferenças entre a utilização de Agentic AI e os outros modelos.

→ Análise do comportamento desviante do DeepSeek V2:

O desempenho anómalo do DeepSeek pode ser explicado por fatores estruturais e operacionais.

O modelo possui menor número de parâmetros e foi otimizado para eficiência e custo, o que reduz a profundidade semântica e a retenção contextual. O treino baseou-se em corpora gerais multilíngues, sem reforço médico, resultando em vocabulário clínico superficial e fraca expansão de tópicos hierárquicos. Além disso, parâmetros de inferência conservadores (temperatura baixa, limites curtos de tokens) e interpretação literal de marcadores ("Respiratório:", "Cardiovascular:") levaram o modelo a produzir respostas demasiado resumidas.

Este desempenho aproximou-se de sistemas de resumo automático, comprimindo informação em vez de a expandir, o que explica a escassez de conteúdo observada.

Resta saber se apenas foi um comportamento isolado ou é uma limitação do modelo, mas tal não faz parte dos objetivos desta Tese.

6.3.5 Controlo e Validação dos Resultados

Não foi pedido a cada modelo, individualmente, para explicar como construiu os textos sintéticos – tal percurso, apesar de extremamente intrigante e desafiador, não faz parte do objetivo desta Tese.

Foi pontualmente aplicada a consistência cruzada com intervenção do autor e iterações suplementares quando necessário.

A apreciação qualitativa já foi referida anteriormente.

As notas finais foram revistas clinicamente por neonatologista (autor), assegurando terminologia adequada e coerência técnica, semântica e temporal.

Seguidamente, foi pedido a um conjunto de três neonatologistas que classificassem os 175 exemplares em três níveis de qualidade:

A) Alta qualidade: indistinguível de um diário clínico real; mantém coerência semântica, cronológica e técnica.
B) Média qualidade: altamente plausíveis, sem erros grosseiros, o tipo de texto que podia ser escrito por um Neonatologista em início de formação.
C) Má qualidade: incompleto, erros grosseiros, quer semânticos, temporais ou técnicos. Inaceitável como texto clínico.

Os classificados com nível C foram removidos do Dataset Final (Figura 47 – Fluxo da depuração do Dataset Sintético).

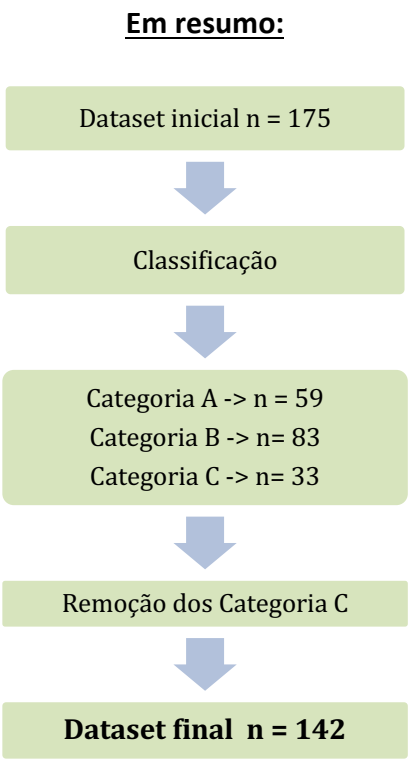


Figura 47 – Fluxo da depuração do Dataset Sintético

6.3.6 Validação do Dataset e Limitações – Análise Crítica

Devemos apontar algumas limitações a esta estratégia:

1) inerentes à geração sintética (ausência de ruído real, dependência dos prompts, enviesamento de linguagem);

2) possível perda de independência (o autor é simultaneamente o tutor dos modelos utilizados e pode influenciar os resultados criados).

Como pontos fortes da estratégia podemos referir:

1) a utilização de modelos diferentes e nenhum dos que vai utilizado para a análise e/ou extração de conteúdo dos diários na próxima fase;

2) a utilização de um conjunto de validadores externos (pares) para depuração e garantia de qualidade;

3) eliminação de todos os textos de qualidade inferior (C);

4) utilização de modelos online state of the art e não as versões offline executáveis em equipamentos domésticos.

Como propostas de melhoria futura sugerimos o ajuste fino com corpus neonatal.

6.3.7 Aplicações

Entre as possíveis aplicações do dataset criado, podemos referir:

- Teste de pipelines de NLP para extração de variáveis clínicas.
- Desenvolvimento de sistemas de apoio à decisão nutricional.
- Avaliação de interoperabilidade (HL7 FHIR©, openEHR©).
- Treino de modelos de mineração de processos para mapear trajetórias de evolução neonatal.

Neste momento, temos o Projeto descrito segundo a metodologia CRISP-MED-DM e um Dataset Sintético validado para utilizar como teste.

7 MODELO LLM A UTILIZAR

Neste capítulo será abordado processo de escolha do LLM mais adequado ao objetivo do projeto.

Como principais critérios de decisão temos, por um lado, o hardware de base que limita os modelos e aplicações utilizáveis, e por outro, o desempenho dos vários LLMs com base em scores objetivos.

7.1 Hardware

De uma maneira geral, podemos afirmar que todas as estratégias do projeto sempre se pautaram por manter um nível acessível de hardware. O objetivo é a integração e execução do fluxo completo no hardware disponível na maioria das UCIN.

Assim, foi utilizado um computador portátil Apple MacBook Air com processador Apple M3, 8 núcleos CPU e 10 núcleos GPU, 16 GB de memória unificada.

O desempenho deste sistema é aproximadamente equivalente a um portátil Windows equipado com processador Intel Core Ultra 7 (ou Core i7-1360P), 16 GB DDR5 RAM e GPU integrada Intel Arc.

Este hardware permitiu a execução de modelos quantizados até ~ 15 milhares de milhão de parâmetros e ~ 9 Gb.

A maioria dos equipamentos disponíveis não terá mais de 16 Gb de memória unificada ou RAM.

7.2 Método de Seleção do LLM

7.2.1 Fluxo Sistematizado de Extração de Variáveis

A escolha do LLM foi planeada de acordo com um processo baseado no seguinte software:

- Sistema operativo macOS Tahoe© 26[10].
- Utilização das aplicações:
  - Gradio© [11] para interação em modo GUI com drag and drop. Apesar de não ser a aplicação principal, funcionou como agregador de entradas e saídas.
  - LM Studio© [12] para execução local de modelos quantizados, em formato GGUF, com suporte para CPU/GPU *offload* e janela de contexto possível até 50 000 tokens sempre com suporte visual (GUI privilegiado sempre). Permite ajustar múltiplos parâmetros do modelo: temperatura, TopK, TopP, limitar a extensão da resposta e até definir *outputs* estruturados (JSON), entre outras possibilidades.
- Ambiente de apoio: Python© [13], bibliotecas transformers, langchain, pandas, matplotlib, openai, etc.

Posteriormente, a comparação dos modelos foi efetuada com métricas reconhecidas, executadas no Orange© [14] e com apoio de Python© para algumas funções.
Assim, procedeu-se a:

- Recolha manual dos dados reais do dataset, para estabelecer a Fonte de Verdade (*Ground Truth);*
- Comparação dos resultados obtidos por cada modelo com a *Ground Truth;*
- Utilização de métricas universalmente reconhecidas para a comparação dos modelos (ver 7.2.2 Métricas Utilizadas para Avaliação dos Modelos);
- Conclusões e escolha do modelo a utilizar na fase seguinte do Projeto.

[10] Apple Inc. (2025). *macOS Tahoe* (versão 26). https://www.apple.com/macos/
[11] Abid, A., Abdalla, A., Agrawal, N., et al. (2019). *Gradio: Hassle-free sharing of machine learning models*. https://gradio.app
[12] LM Studio AI. (2025). *LM Studio* (versão 0.3.31). https://lmstudio.ai
[13] Python Software Foundation. (2025). *Python* (versão 3.13). https://www.python.org
[14] Orange Data Mining. (2025). *Orange* (versão 3.3.8). https://orangedatamining.com
Na realização deste Projeto foram sendo atualizadas para as versões mais recentes disponíveis, sendo a última atualização e acesso efetuados até 30 de novembro de 2025.

O fluxo de análise é apresentado abaixo (Figura 48 – Esquema de funcionamento para extração de variáveis a partir de texto clínico).

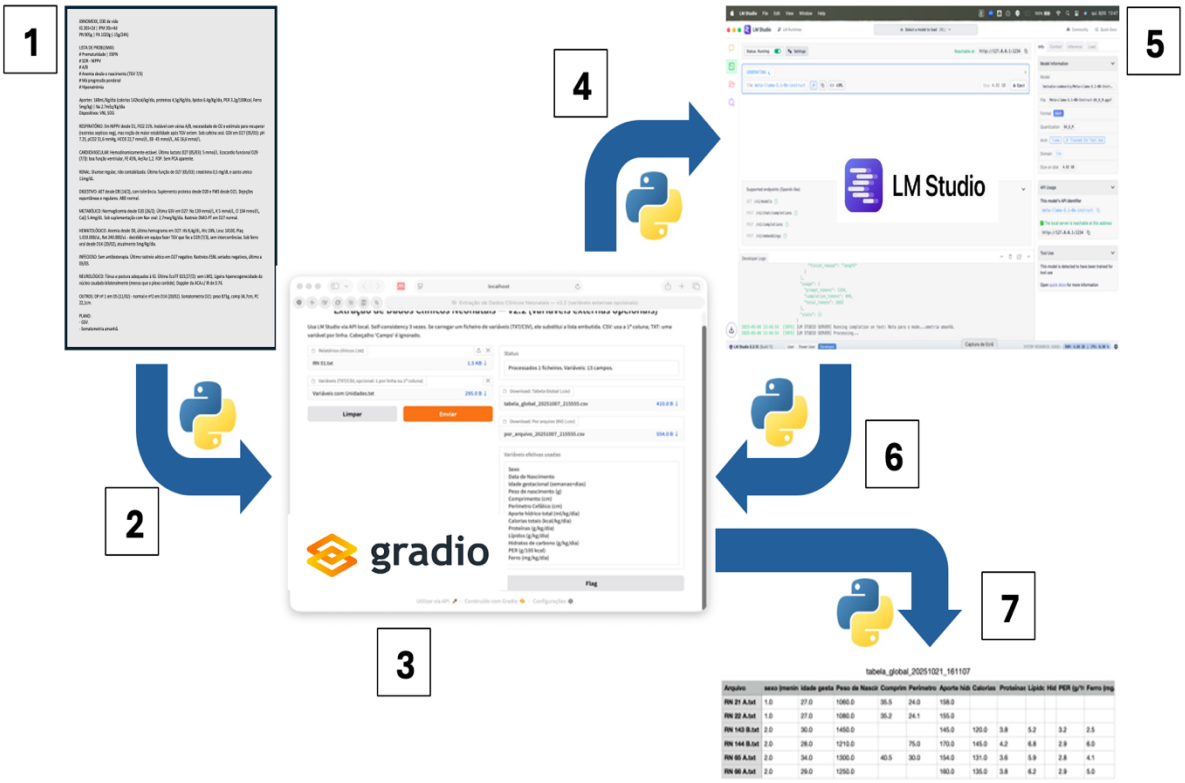


Figura 48 – Esquema de funcionamento para extração de variáveis a partir de texto clínico

Globalmente, o fluxo processa-se assim:

a) os 142 textos do dataset utilizado são formatados de acordo com o modelo "RN ###.txt".
b) execução de app em Python e lançamento do Gradio.
c) esta interface permitiu a interação baseada num ambiente gráfico *drag and drop*, bem como o acesso aos dados extraídos.
    - nesta interface é possível carregar um ficheiro .txt com a lista das variáveis a recolher.
    - os textos clínicos (formato .txt) são selecionados e carregados também aqui.
    - mais tarde, os resultados finais são apresentados também aqui, num ficheiro .csv para download. Este contém a identificação de cada texto (nome genérico RN ###.txt) e os valores recolhidos para cada variável.

- permite ainda ter acesso ao tempo decorrido para cada grupo de ficheiros txt em análise.

4) cada "RN ###.txt" é enviado para o LMStudio através da app Python.
5) previamente, foi efetuado o upload de cada modelo no LMStudio, aplicando-se parâmetros uniformes e standardizados em todos os testes:
    - Temperatura = 0,15 - valor reduzido para restringir a aleatoriedade da geração textual, favorecendo respostas determinísticas e realistas.
    - Top-K = 5 - limita a amostragem aos cinco tokens de maior probabilidade, adequado à extração de dados objetivos predominantemente numéricos. A única variável categórica (sexo) é binária, sem grande amplitude lexical.
    - Top-P = 0,95 - foi mantido o valor habitual.
6) o *Output* do LLM é enviado para o Gradio.
7) o .csv com a tabela global dos resultados é gerado e disponibilizado para download.

Este processo foi executado para cada um dos modelos avaliados, sendo anotado o tempo total de execução da análise dos 142 textos.

O LMStudio permitiu ajustar algumas salvaguardas que evitam a execução de modelos com tamanho que possam bloquear e congelar o sistema.

7.2.2 Métricas Utilizadas para Avaliação dos Modelos

A avaliação foi efetuada com a utilização de métricas universalmente reconhecidas para a comparação dos modelos.

A avaliação do desempenho dos modelos foi realizada com recurso às métricas de classificação exatidão, precisão, sensibilidade e F1-score; e às métricas de erro MAE, RMSE e MAPE [102], [103]. As definições e aplicabilidade de cada uma encontram-se nas Tabela 4 e Tabela 5.

Tabela 4 – Métricas de Classificação

Métrica	Definição	Fórmula
Accuracy (Exactidão)	Proporção de previsões corretas em relação ao total de previsões. Mede o desempenho global do modelo.	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision (Precisão)	Proporção de previsões positivas corretas em relação ao total de previsões positivas realizadas.	$\frac{TP}{TP + FP}$
Recall (Sensibilidade)	Proporção de casos positivos corretamente identificados em relação ao total real de casos positivos.	$\frac{TP}{TP + FN}$
F1-Score	Média harmónica entre Precision e Recall. Mede o equilíbrio entre falsos positivos e falsos negativos.	$\frac{Precision \cdot Recall}{Precision + Recall}$
Taxa de Falsos Positivos (FP)	Proporção de casos negativos incorretamente classificados como positivos.	$\frac{FP}{FP + TN}$
Taxa de Falsos Negativos (FN)	Proporção de casos positivos incorretamente classificados como negativos.	$\frac{FN}{FN + TP}$

TP = Verdadeiros Positivos; TN = Verdadeiros Negativos; FP = Falsos Positivos; FN = Falsos Negativos.

Tabela 5 – Métricas de Erro

Métrica	Definição	Fórmula
MAE (Mean Absolute Error)	Média dos erros absolutos. Mede o erro médio em unidades reais.	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - p_i\rvert$
RMSE (Root Mean Squared Error)	Raiz quadrada da média dos erros ao quadrado. Penaliza mais erros elevados.	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - p_i)^2}$
MAPE (Mean Absolute Percentage Error)	Erro médio absoluto em percentagem relativamente ao valor real.	$\frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i - p_i}{y_i}\right\rvert$

y_i = valor real; p_i= valor previsto; n = número de observações

7.2.3 Modelos Considerados para Análise

De acordo com as limitações do hardware previamente referido, foram selecionados os modelos compatíveis no LMStudio – ver Tabela 6:

Tabela 6 – Modelos testados (parâmetros, tipo de quantização, tamanho e memória "real" ocupada com o funcionamento de todas as aplicações); B = milhares de milhão de parâmetros.

Modelo	Parâmetros	Quantização	Tamanho	Memória Ocupada
Gemma 3	12 B	Q4_K_M	8,15 Gb	14,4 Gb
Qwen 3	4 B	4bit	2,28 Gb	11 Gb
Granite-4.0-h-tiny	7 B	Q4_K_M	4,23 Gb	13 Gb
Meta-Llama-3.1 -Instruct	8 B	Q4_K_M-8bit	4,9 Gb	13,6 Gb
Microsoft.phi-4-reasoning-plus	14,7 B	4bit	8,26 Gb	14,5 Gb
Qwen3-claude-sonnet-4-reasoning-distill	8 B	Q4_K_M	5,03 Gb	13 Gb

Os seis modelos testados correspondem aos disponíveis para utilização no LMStudio.

7.2.4 Resultados – Extração de Variáveis do Dataset

a) Tempo para processar textos clínicos

A escolha do modelo implica também a avaliação do tempo total para processar todos os 142 textos e extrair as variáveis pedidas. Esta avaliação está na Tabela 7.

Tabela 7 – Tempo de processamento e características dos modelos

Modelo	Parâmetros	Quantização	Tempo Total
Gemma 3	12 B	Q4_K_M	4 h 05 min
Qwen 3	4 B	4bit	1 h 35 min
Granite-4.0-h-tiny	7 B	Q4_K_M	1 h 15 min
Meta-Llama-3.1 -Instruct	8 B	Q4_K_M-8bit	3 h 35 min
Microsoft.phi-4-reasoning-plus	14,7 B	4bit	1h 40 min
Qwen3-claude-sonnet-4-reasoning-distill	8 B	Q4_K_M	5 h 42 min

Objetivamente, existe alguma heterogeneidade entre o tempo total de processamento e as características dos modelos, não sendo possível detetar um padrão consistente – ver Figura 49.

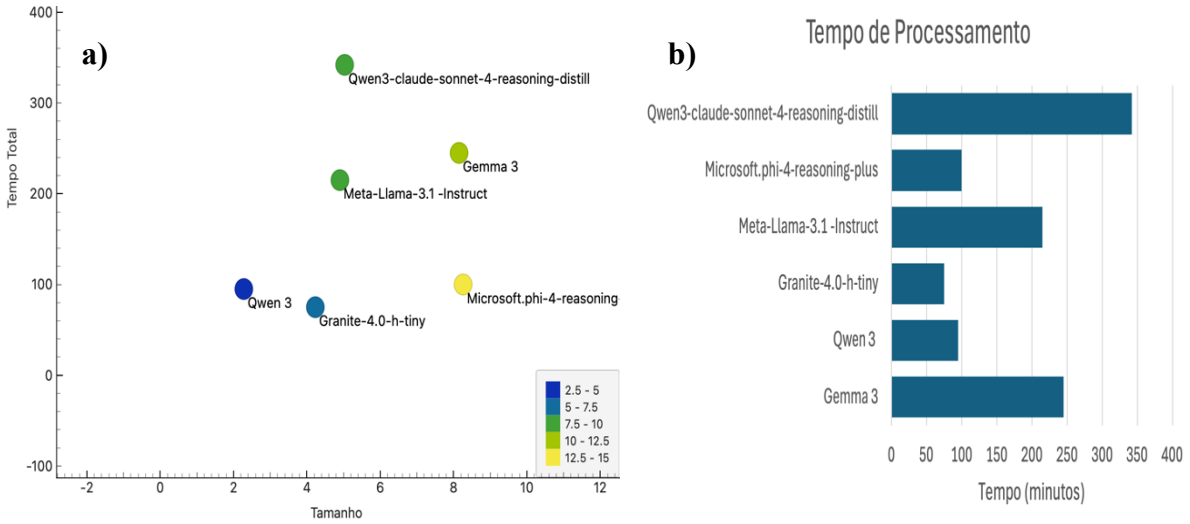


Figura 49 - Comparação do tempo total para processar os 142 textos clínicos e extrair as variáveis pedidas. **a)** apresentam-se as relações entre tempo total (min; eixo y), tamanho (Gb; eixo x) e o número de parâmetros (Dimensão parâmetros em milhares de milhão; cores). **b)** Tempo total de processamento para comparação.

Nesta comparação, verificamos uma grande heterogeneidade no tempo necessário para cada modelo. As correlações de Spearman (**ρs**) observadas foram:

Correlações	**Valor**	**Comentário**
ρs(Tamanho,N.º de Parâmetros)	**+0,977**	Correlação positiva forte, de acordo com o esperado
ρs(Tamanho,Tempo Total)	**+0,211**	Correlação positiva fraca; nem sempre um tamanho maior corresponde a mais tempo
ρs(Parâmetros,Tempo Total)	**+0,087**	Correlação muito fraca entre o número de parâmetros dos modelos e o tempo que necessitaram

Globalmente, podemos verificar que existe uma correlação positiva forte entre o tamanho do modelo (em Gb) e o número de parâmetros (B / milhares de milhões) o que é esperado.

Existe também uma correlação positiva fraca entre o tempo total de processamento dos textos (em horas e minutos) e o tamanho dos modelos (em Gb), o que se traduziu por modelos maiores como o *Microsoft.phi-4-reasoning-plus* demorar bastante menos tempo que *Gemma 3* para tamanhos equivalentes.

E, por fim, uma correlação positiva muito fraca (<0,1) entre o número de parâmetros e o tempo total: *Qwen3-claude-sonnet-4-reasoning-distill* foi o mais lento, enquanto o *Meta-Llama-3.1 -Instruct* para número semelhante de parâmetros demorou menos tempo ou em comparação com *Microsoft.phi-4-reasoning-plus*, mais rápido e com mais parâmetros.

Em resumo, poucas conclusões com impacto no Projeto se retiram destes dados. Necessitamos de métricas mais diferenciadas, já referenciadas anteriormente (ver 7.2.2. - Métricas utilizadas para avaliação dos modelos).

b) Desempenho de cada modelo na extração de variáveis categóricas

Neste dataset apenas existe uma variável categórica (Sexo M / F). Para avaliar o desempenho de cada modelo, foram consideradas as métricas e os resultados obtidos, que constam da Tabela 8 – Scores para variáveis categóricas.

Nesta tabela, não é possível distinguir quando o modelo não extrai/ignora a variável ou apresenta um valor errado.

Tabela 8 – Scores para variáveis categóricas.

modelo	TP	TN	FP	FN	n	Accuracy	Precision	Recall	F1
a_gemma3_12b	57	85	0	0	142	1.000	1.000	1.000	1.000
b_granite_7b	56	85	0	1	142	0.993	1.000	0.982	0.991
c_llama_3_8b	33	85	0	24	142	0.831	1.000	0.579	0.733
d_phi4_14b	12	85	0	45	142	0.683	1.000	0.211	0.348
e_qwen3_4b	57	85	0	0	142	1.000	1.000	1.000	1.000
f_Qwen3_claude_8b	24	85	0	33	142	0.768	1.000	0.421	0.593

Para isso, os valores em falta foram convertidos em FN para cada uma das classes M e F, e os scores recalculados, para cada uma das classes M e F, como consta na Tabela 9 a) e b) – Scores das variáveis categóricas. Em a), calculados para classe M e em b), calculados para classe F. Em cada classe, os valores em falta foram contabilizados como FN (Falso Negativo).

Tabela 9 a) e b) – Scores das variáveis categóricas. Em a), calculados para classe M e em b), calculados para classe F. Em cada classe, os valores em falta foram contabilizados como FN (Falso Negativo).

a)

modelo	TP	TN	FP	FN	n	Accuracy	Precision	Recall	F1
a_gemma3_12b	57	85	0	0	142	1.000	1.000	1.000	1.000
b_granite_7b	56	85	0	1	142	0.993	1.000	0.982	0.991
c_llama_3_8b	33	85	0	24	142	0.831	1.000	0.579	0.733
d_phi4_14b	12	85	0	45	142	0.683	1.000	0.211	0.348
e_qwen3_4b	57	85	0	0	142	1.000	1.000	1.000	1.000
f_Qwen3_claude_8b	24	85	0	33	142	0.768	1.000	0.421	0.593

b)

modelo	TP	TN	FP	FN	n	Accuracy	Precision	Recall	F1
a_gemma3_12b	85	57	0	0	142	1.000	1.000	1.000	1.000
b_granite_7b	84	57	0	1	142	0.993	1.000	0.988	0.994
c_llama_3_8b	44	57	0	41	142	0.711	1.000	0.518	0.682
d_phi4_14b	25	57	0	60	142	0.577	1.000	0.294	0.455
e_qwen3_4b	83	57	0	2	142	0.986	1.000	0.976	0.988
f_Qwen3_claude_8b	34	57	0	51	142	0.641	1.000	0.400	0.571

Ou seja, pretende-se não apenas salientar quando o modelo identifica e extrai corretamente o valor da variável, neste caso, classe M ou F, mas também apresentar o cenário quando este ignora o item (valores em falta). Para isso, os valores em falta foram contabilizados como FN para cada uma das classes.

Apesar de tudo, os modelos não "trocaram" classes, ou seja, não ocorreram casos em que um M foi erradamente classificado como F e vice-versa.

Nas tabelas apresentadas, o desempenho de um dos modelos (Gemma 3) continua a destacar-se, com valores máximos em todos os scores, quer no cenário a) ou b).

c) Desempenho dos modelos na extração de variáveis contínuas

Neste dataset existem múltiplas variáveis contínuas. Para avaliar o desempenho de cada modelo, foram consideradas as métricas já referidas anteriormente (ver Tabela 5). Os resultados obtidos encontram-se descritos na Tabela 10.

Tabela 10 – MAE, RMSE, MAPE, Missed e Missed Rate (%) dos modelos testados.

Modelo	MAE	RMSE	MAPE	Missed	Miss Rate %
a_gemma3_12b	**0.06 [0,032; 0,096]**	**0.627 [0,414; 0,812]**	**0.082**	**1**	**0.071**
b_granite_7b	15.156 [7,193; 23,688]	138.893 [92,257; 178,904]	40.837	385	27.52
c_llama_3_8b	1.579 [0,121; 4,384]	38.167 [0,849; 66,075]	4.251	638	45.604
d_phi4_14b	5.118 [0,038; 15,236]	87.383 [0,361; 151,349]	13.2	1095	78.27
e_qwen3_4b	8.504 [5,524; 11,791]	61.746 [46,962; 75,925]	1.053	16	1.144
f_Qwen3_claude_8b	0.86 [0,093; 2,223]	12.914 [0,566; 22,223]	0.52	840	60.043

Ao analisarmos a tabela, verificamos que há um modelo (Gemma 3) que se destaca amplamente de todos, ao conseguir extrair 92,9% dos dados, com MAE de 0,06, RMSE de 0,627 e MAPE 0,082 %. Esta diferença é mais evidente se analisarmos a Figura 50.

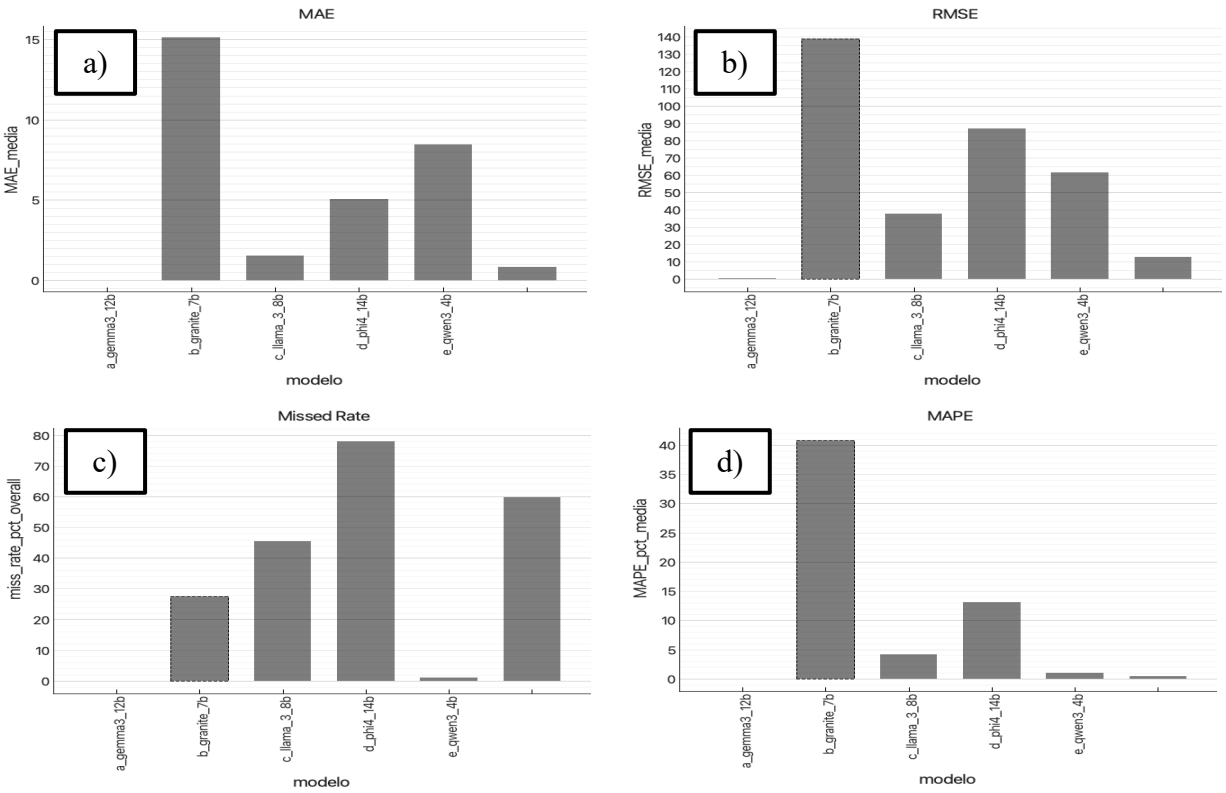


Figura 50 – Comparação entre os modelos: a) MAE; b) RMSE; c) Missed Rate %; d) MAPE.

Modelo	Parâmetros	Quantização	Tamanho	Memória Ocupada	Tempo Total	TP	TN	FP	FN	n	Accuracy	Precision	Recall	F1	MAE	RMSE	MAPE	Missed	Missed Rate %
a_gemma3_12b	12 B	Q4_K_M	8.15 Gb	14.4 Gb	4h 05min	57	85	0	0	142	1.0	1.0	1.0	1.0	0.06 [0.032; 0.096]	0.627 [0.414; 0.812]	0.082	1	0.071
b_granite_7b	7 B	Q4_K_M	4.23 Gb	13.0 Gb	1h 15min	56	85	0	1	142	0.993	1.0	0.982	0.991	15.156 [7.193; 23.688]	138.893 [92.257; 178.904]	40.837	385	27.52
c_llama_3_8b	8 B	Q4_K_M-8bit	4.9 Gb	13.6 Gb	3h 35min	33	85	0	24	142	0.831	1.0	0.579	0.733	1.579 [0.121; 4.384]	38.167 [0.849; 66.075]	4.251	638	45.604
d_phi4_14b	14.7 B	4bit	8.26 Gb	14.5 Gb	1h 40min	12	85	0	45	142	0.683	1.0	0.211	0.348	5.118 [0.038; 15.236]	87.383 [0.361; 151.349]	13.2	1095	78.27
e_qwen3_4b	4 B	4bit	2.28 Gb	11.0 Gb	1h 35min	57	85	0	0	142	1.0	1.0	1.0	1.0	8.504 [5.524; 11.791]	61.746 [46.962; 75.925]	1.053	16	1.144
f_Qwen3_claude_8b	8 B	Q4_K_M	5.03 Gb	13.0 Gb	5h 42min	24	85	0	33	142	0.768	1.0	0.421	0.593	0.86 [0.093; 2.223]	12.914 [0.566; 22.223]	0.52	840	60.043

Tabela - Resultados globais da análise ao desempenho dos 6 modelos considerados.
Parâmetros em milhares de milhões (B); Quantização referida na descrição dos modelos, nem sempre equivalente; Tamanho em Gb referido na descrição do modelo; Memória ocupada total com o modelo e as aplicações necessárias (em Gb); Tempo total em horas e minutos para o processamento completo dos 142 textos; TP/TN/FP/FN expressos em n; n = número de textos processados (total = 142); Accuracy, Precision e F1 variando 0-1; MAE - Erro Absoluto Médio; RMSE - Raiz do Erro Quadrático Médio; MAPE - Erro Percentual Absoluto Médio; Missed – número de itens não identificados e extraídos; Missed Rate (em %).

7.3 Comentários ao Desempenho dos Modelos

Após o processo descrito em 7.2 – Método de Seleção do LLM, podemos concluir que os seis modelos em análise tiveram desempenhos muito diferentes.

Na globalidade, podemos afirmar que, nesta amostra e neste conjunto de tarefas, o número de parâmetros não foi o fator mais decisivo, pois o modelo Phi 4 com 14,7 B parâmetros teve o pior desempenho, em clara oposição ao melhor desempenho entre todos do Gemma 3 com 12 B parâmetros. Mesmo entre quantidade de parâmetros equivalente – 7 a 8 B parâmetros – os modelos Llama 3, Granite e Qwen 3 Claude tiveram resultados muito diferentes.

Se o objetivo fosse a rapidez de processamento, o Granite 7B seria o melhor, mas com uma das piores pontuações em termos de qualidade.

Atendendo ao tipo de tarefa considerada nesta avaliação, a qualidade dos dados extraídos é o aspeto mais importante – trata-se de dados médicos, para utilização posterior em contexto clínico. E neste item, o Gemma 3 - 12 B tem um desempenho quase perfeito. Se tivermos de sacrificar qualidade por um pouco mais de velocidade (neste Setup, é certo), o Qwen 3 – 4 B seria uma boa opção. Mas implica perda de qualidade, o que é inaceitável e perigoso em contexto clínico.

Apesar de se tratar de um dataset sintético, cuja conceção foi já abordada (ver 6.3 – Dataset Sintético), este reproduz a população habitual numa UCIN e, por isso, contém algumas assimetrias reais, como por exemplo, uma proporção entre M e F diferente de 50:50. Na avaliação dos modelos, podemos constatar que estes não apresentaram enviesamento dos scores.

Uma análise mais aprofundada às diferenças encontradas na avaliação do desempenho destes seis modelos seria complexa e interessante, com certeza um assunto a revisitar em trabalho futuro, mas fora do âmbito desta Tese.

Em resumo, o modelo com maior qualidade na extração de dados clínicos foi o **Gemma 3 – 12 B** e demoraria cerca de 1 min e 30 s a processar um único texto. Por isso, foi o modelo escolhido para a próxima parte do projeto.

8 MODELAÇÃO E AVALIAÇÃO DE RESULTADOS

Recordando o Objetivo do Projeto e a Questão de Investigação:

"Qual a capacidade de um LLM para elaborar uma proposta de ajuste nutricional com base em diários clínicos?"

8.1.1 Descrição do Processo do Ponto de Vista Humano

O ponto de partida será a descrição do processo "manual", tal como executada por um médico no seu dia a dia (ver Figura 51 – Processo Simplificado de Ajuste de Aportes Nutricionais):

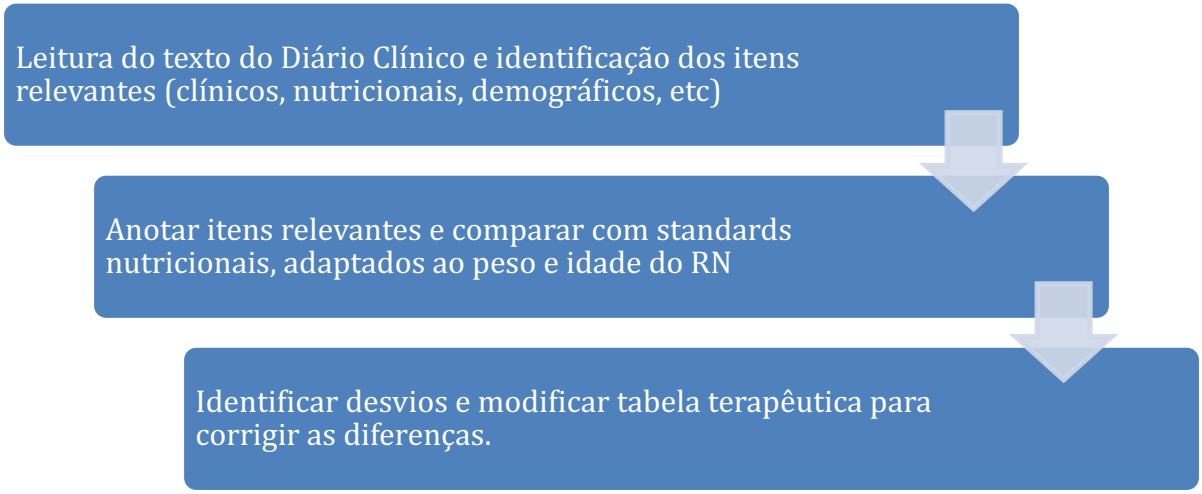


Figura 51 – Processo Simplificado de Ajuste de Aportes Nutricionais

Este processo apresenta várias limitações, podendo ainda ser facilmente indutor de erros e sujeito a vieses múltiplos – problemas amplamente descritos na literatura e que ultrapassam o âmbito desta Tese. Estes podem ainda ser agravados pelo cansaço e pela repetibilidade da tarefa, e ainda amplificados pela constante divisão da atenção numa UCIN com vários RN.

8.1.2 Organização do Processo Automatizado com IA

O processo deve, de certo modo, replicar o fundamental do apresentado no ponto anterior. Deve analisar um texto clínico de forma automatizada e autónoma, extrair os itens necessários para estabelecer o estado atual dos aportes nutricionais, depois deve comparar com as metas nutricionais individualizadas e ajustadas ao RN em questão. Posteriormente, deve identificar desvios e finalmente elaborar uma proposta de ajuste nutricional.

Genericamente, utilizou-se uma interface gráfica (Gradio©) baseada em Python© que permitia a integração do LMStudio ©, dos referenciais (Z-Scores e Standards Nutricionais) e dos textos clínicos.

Ora esta abordagem passou por várias fases de desenvolvimento, com múltiplos ciclos de tentativa e erro.

Globalmente, as maiores dificuldades estiveram relacionadas com as limitações inerentes à utilização integral do LLM. Após a extração dos dados, com uma metodologia decalcada da descrita anteriormente (ver 7.2 Método de Seleção do LLM) e, como tal, já validada; ocorreram múltiplas dificuldades na utilização efetiva dos dados.

Resumidamente, para que o processo fosse integralmente realizado pelo LLM, todos os elementos deveriam ser incluídos na prompt fornecida. Esta opção implica que os referenciais para classificação de zscores e standards nutricionais por sexo, grupo de peso do RN e idade gestacional fossem carregados na prompt. Ora rapidamente as prompts excederam os 150 000 tokens, o que ficava muito para além da capacidade quer do modelo (139 000 tokens), quer do hardware utilizado.

Estas limitações condicionaram a opção por uma estratégia diferente.

8.2 Processo Detalhado

A metodologia foi reformulada, optando-se por dividir o Processo Integral em diferentes Tarefas, cada uma delas criada sequencialmente após garantir que a anterior funcionava plenamente.

Após algumas iterações, verificou-se que o método de registo e partilha de dados entre as diferentes tarefas teria de ser rigidamente estruturado, para o qual se utilizou o formato JSON.

Seguidamente, apresentamos na Figura 52 o esquema global do Processo utilizado.

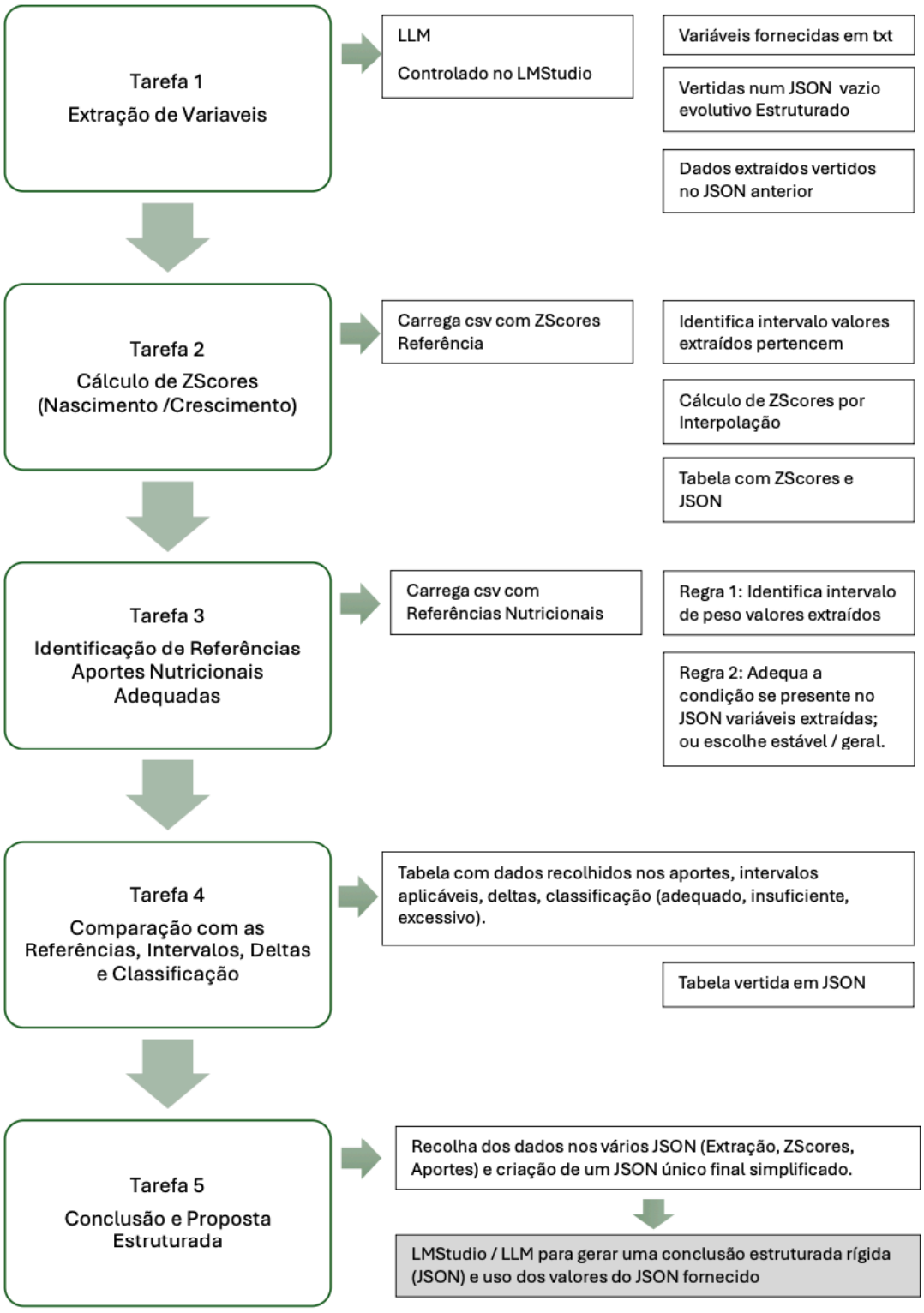


Figura 52 – Diagrama do Processo Global

8.2.1 Tarefa 1 - Extração da Informação Clínica Relevante

Esta tarefa foi decalcada do método já referido antes (ver 7.2.1 Fluxo Sistematizado de Extração de Variáveis). Relembrando, o processo é baseado em Python©, que conjuga as várias tarefas, com o interface gráfico drag and drop do Gradio©. Neste GUI, são selecionadas: o texto clínico para análise (.txt); o conjunto de variáveis a extrair (.txt); e a ligação ao LMStudio – ver Figura 53.

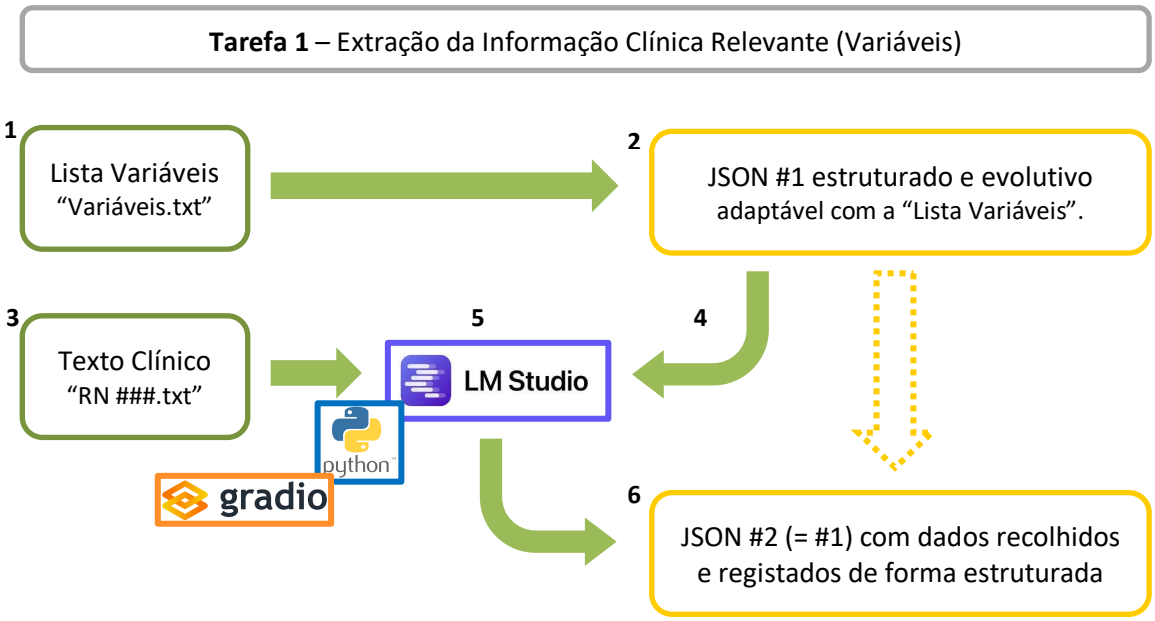


Figura 53 – Diagrama da Tarefa 1

A **Tarefa 1** inicia-se com a introdução da lista de variáveis (1) num .txt carregado através Gradio©; esta lista é transposta para um JSON estruturado, vazio (2). De seguida, é selecionado (3) um texto clínico - proveniente do dataset sintético já utilizado antes (ver 6.3 Dataset Sintético). É gerada uma prompt para o LLM contendo o JSON "vazio" (4) e o texto clínico, sendo ambos enviados para o LLM, parametrizado integralmente no LMStudio. Neste passo (5) ocorre a extração de variáveis pelo LLM e os dados colhidos vertidos no JSON fornecido, sendo gerado um JSON final pelo LLM com os dados sistematizados e estruturados (6). Este JSON vai ser utilizado posteriormente nas Tarefas seguintes.

Destaca-se a possibilidade de a Lista de Variáveis poder evoluir com mais ou menos variáveis a serem extraídas. A utilização de um JSON estruturado criado na

sequência da lista de variáveis permite manter uma estrutura coerente (aceita boolean, texto, float) e em última análise obriga a unidades do sistema LOINC.

Esta estratégia resolveu vários problemas:

a) nas tentativas iniciais, a prompt era sempre modificada de cada vez que se mudavam as variáveis (mesmo uma pequena alteração reiniciava todo o código Python).

b) o registo e exportação dos dados, inicialmente em .csv, era passível de alguma criatividade por parte do LLM.

c) assim, apesar do LLM trabalhar com os mesmos parâmetros utilizados antes (temperatura 0,15; top-K = 5 e top-p = 0,95), "obriga" a registar e exportar os dados num formato e estrutura rígidos.

d) este formato de registo vai usar sempre as mesmas unidades e estrutura ao longo das várias tarefas (ex: peso de nascimento vai ser sempre referenciado da mesma maneira).

Na Figura 54 encontram-se alguns instantâneos da Tarefa 1.

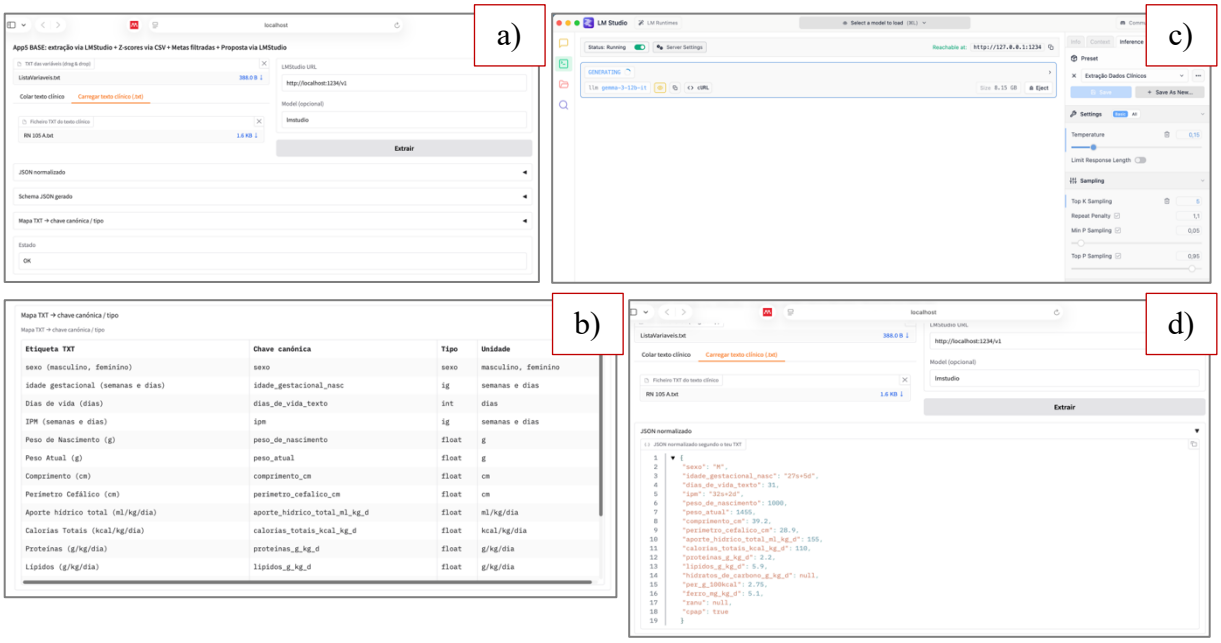


Figura 54 – **Tarefa 1**: a) carregar "lista de variáveis.txt" contendo as variáveis a extrair e o texto clínico a analisar "RN ##.txt"; b) conversão do texto simples da lista de variáveis para chave canónica (gera o JSON #1 vazio); c) LMStudio corre o LLM com os parâmetros ajustados; d) variáveis extraídas e vertidas em formato normalizado no JSON #2.

8.2.2 Tarefa 2 – Cálculo de Z-Scores

Nesta tarefa, pretende-se calcular o Z-Score de variáveis somatométricas, como o peso de nascimento, peso atual, comprimento e perímetro cefálico (ver Figura 55). Os dados originais extraídos do texto clínico estão registados no JSON #2 em unidades sistema LOINC. As idades são convertidas de "text" para "float" (exemplo: 26s+6d em 26,857 semanas).

Paralelamente, é carregado um .csv com os Z-Scores de referência por sexo, idade gestacional de acordo com os de Fenton e Intergrowth-21, com Z-Scores -3; -2; -1; 0; 1; 2; 3. Com base nestes valores Z, são calculados os Z exatos através de interpolação contínua. O .csv é editável e evolutivo, podendo ser modificado se forem publicados outros referenciais.

Após estes cálculos, executados em Python, os resultados são apresentados numa tabela na interface do Gradio e também exportados para um JSON #3, para utilização futura (Figura 55).

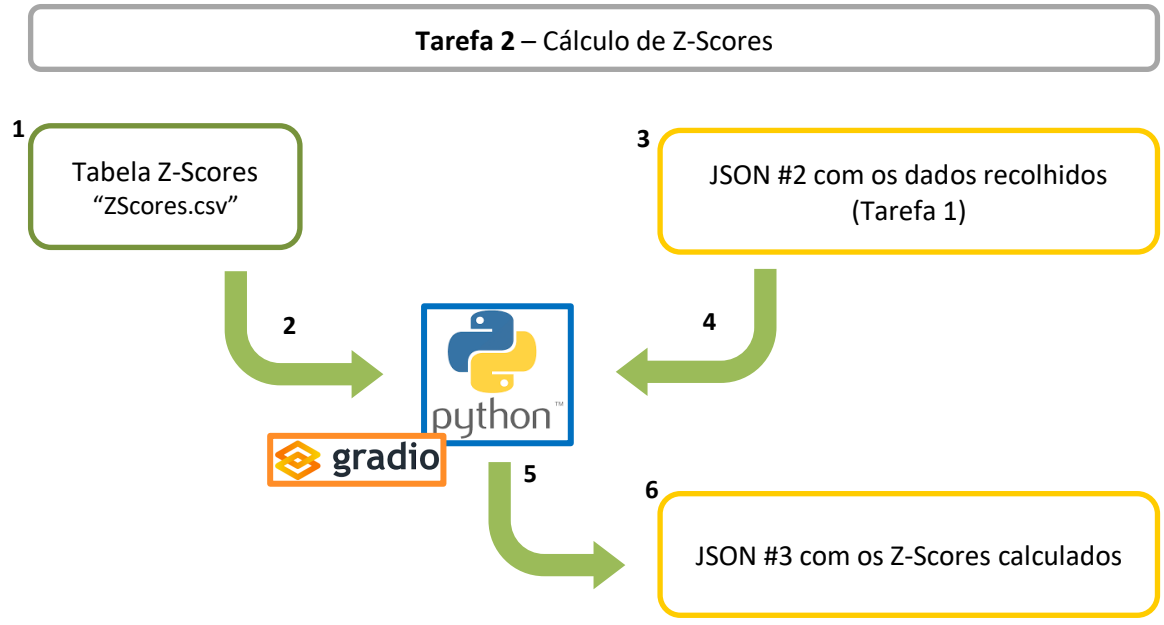


Figura 55 – Diagrama da Tarefa 2: 1-2) o referencial de Z-Scores (-3 a 3 para cada variável somatométrica, idade gestacional e sexo) em .csv é carregado através do interface do Gradio; 3-4) o JSON #2 com os dados extraídos do texto clínico é utilizado para calcular o Z-Score para cada variável somatométrica (valores obtidos por interpolação com base no referencial em .csv); 5-6) os Z-Scores calculados são vertidos no JSON #3 para utilização posterior.

Na Figura 56 encontram-se alguns instantâneos da Tarefa 2.

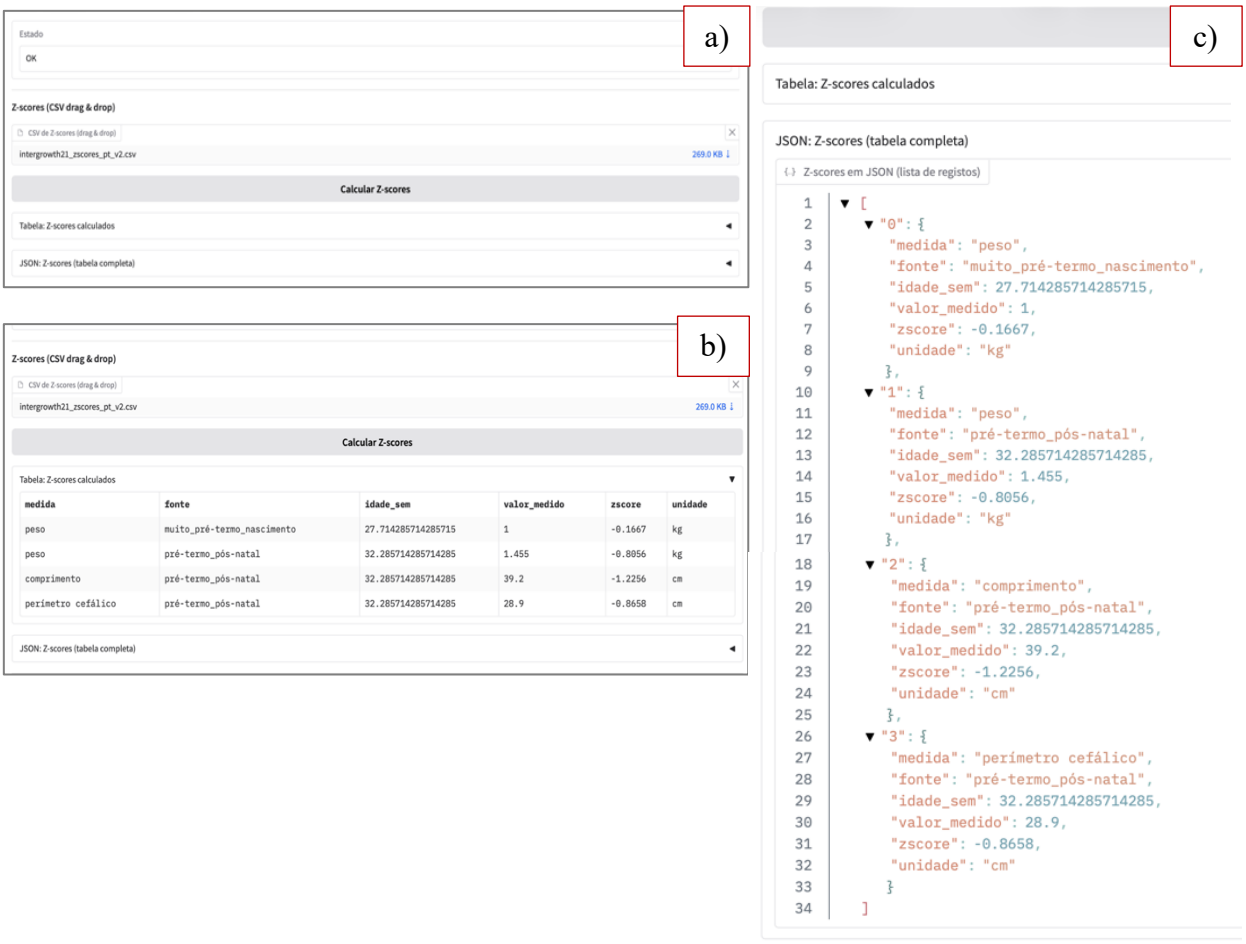


Figura 56 – **Tarefa 2**: a) carregar referências "Z-Scores.csv"; b) tabela com Z-Scores calculados para as variáveis extraídas (dados contidos no JSON #2); c) JSON #3 com os Z-Scores calculados.

8.2.3 Tarefa 3 - Identificação das Referências Nutricionais adequadas ao RN

A Tarefa 3 faz corresponder o nosso RN, com as características clínicas extraídas do texto do diário clínico, ao conjunto de metas de aportes nutricionais (ver Figura 57).

Simplificando, o RN pode ser classificado em termos de peso de nascimento (<1000 g; 1000 a <1500 g; 1500 a <2500 g; 2500 g ou superior) e condição clínica (estável; presença de patologia como DBP ou PCA). Estes dados estão já presentes no JSON #2.

Assim, são selecionados os intervalos adequados para cada item dos aportes nutricionais (AHT, calorias, proteínas, etc). Estes são apresentados numa tabela na interface do Gradio e vertidos num JSON #4.

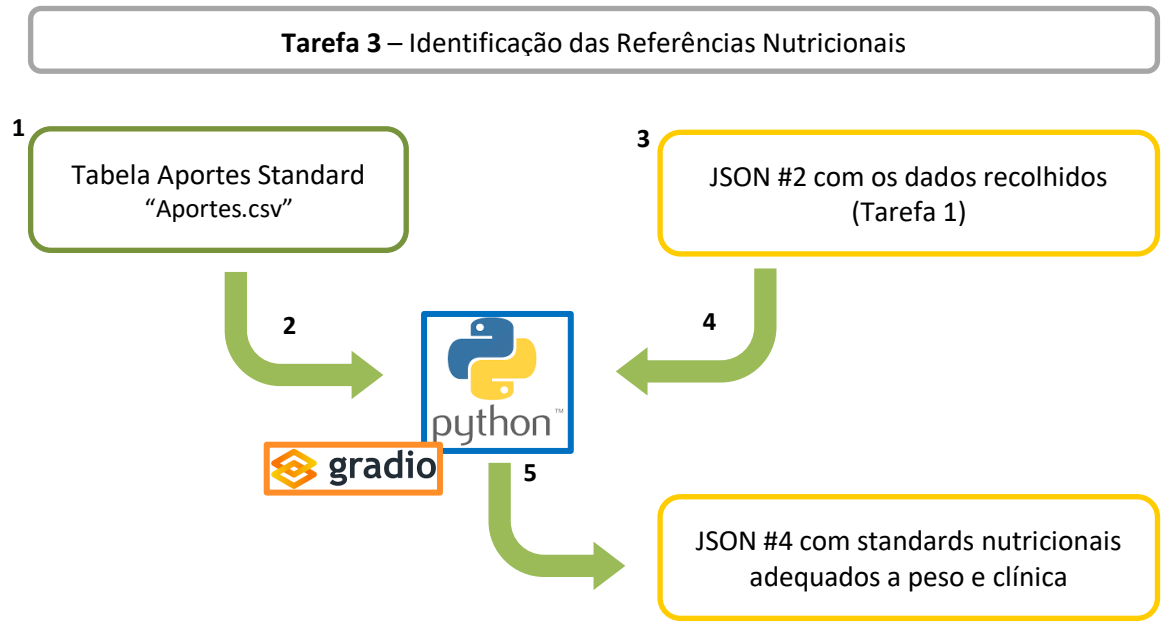


Figura 57 – Diagrama da Tarefa 3

Nesta tarefa, é carregado através do Gradio o conjunto de referências para os aportes nutricionais, subdivididos em grupos de peso e características clínicas (1-2).

Este é utilizado com os dados extraídos do diário clínico com o JSON #2 (3-4) para encontrar os valores adequados (standard) para o RN individualizado, tendo em conta as características já referidas.

Este novo referencial individualizado é visualizado numa tabela do Gradio e vertido num JSON #4 para utilização posterior.

Na Figura 58 encontram-se alguns instantâneos da Tarefa 3.

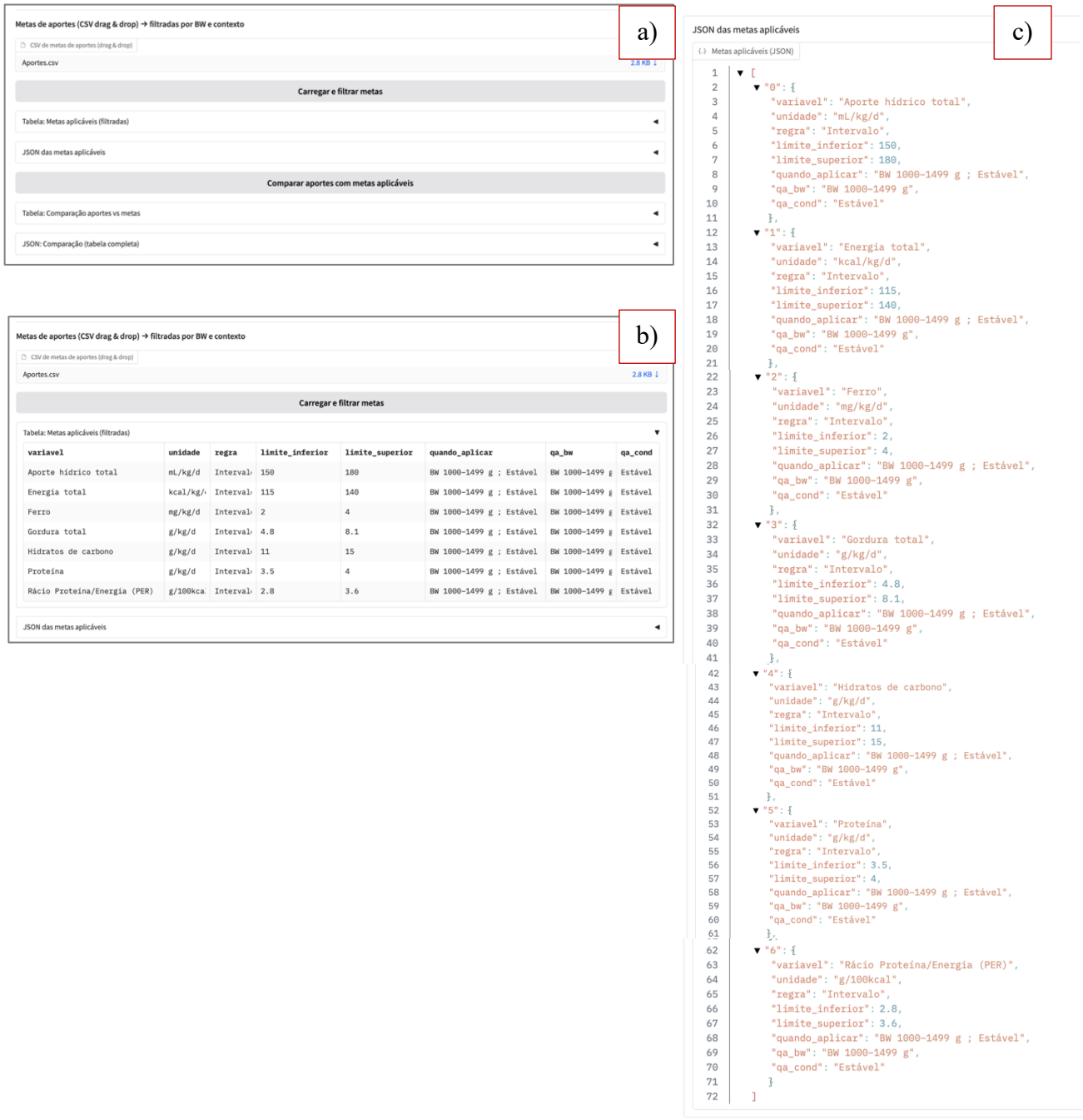


Figura 58 – **Tarefa 3**: a) carregar referências para aportes nutricionais "aportes.csv"; b) tabela com os intervalos de aportes estandardizados para as características do RN; c) JSON #4 com os intervalos admissíveis para o RN em análise.

8.2.4 Tarefa 4 - Comparação com os Standards Adequados e Identificação de Ajustes para Correção

Nesta tarefa é suposto comparar-se a adequação entre os aportes atuais e as metas de referência; classificar como "adequado", "inferior ao adequado", "superior ao adequado" e finalmente apresentar o diferencial de valores. Os valores "atuais" são fornecidos pelo JSON #2.

A comparação, classificação e diferencial são apresentados numa tabela na interface do Gradio e são exportados num JSON #5 – ver Figura 59.

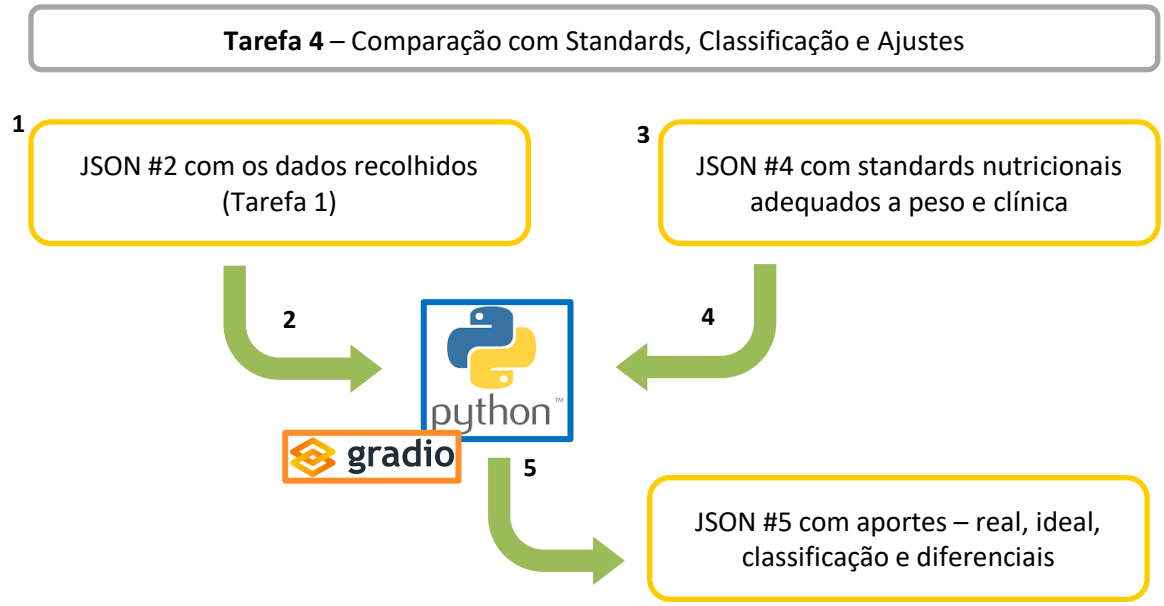


Figura 59 – Diagrama da Tarefa 4

A Tarefa 4 inicia-se com a leitura (1-2) do JSON #2 que contém os valores dos aportes nutricionais atuais do RN (extraídos do diário clínico) e do JSON #4 que contém os intervalos ideais para os mesmos aportes (3-4), adequados ao RN individualmente de acordo com os standards.

Posteriormente, é realizada uma comparação entre os dois, classificando como adequado, excessivo ou insuficiente (Python) e extraída uma tabela que é apresentada no Gradio.

Esta tabela é depois vertida num JSON #5 e utilizada na conclusão.

Na Figura 60 encontram-se alguns instantâneos da Tarefa 4.

Figura 60 – **Tarefa 4**:
a) Tabela com a comparação entre os valores medidos e os intervalos ideais; classificação e diferenciais.
b) JSON #5 com os elementos da tabela anterior.

Este JSON #5 concentra praticamente toda a informação relevante do processo, ao conter os valores reais extraídos do diário clínico e a adequação dos mesmos.

8.2.5 Tarefa 5 – Conclusão e Elaboração de uma Proposta Estruturada

Por fim, na Tarefa 5 pretende-se englobar todos os passos anteriores e resumir numa conclusão e proposta de ajuste.

De uma maneira geral, queremos informar quais os dados recolhidos, como se ajustam ao RN em particular e, necessitando de ajuste, qual o diferencial. Esta conclusão deve ser em texto natural, pelo que se recorre ao LLM através do LMStudio.

Para que o LLM gere um texto relativamente conciso, em linguagem médica, e referindo todos os pontos fundamentais, foi necessário implementar limites rígidos.

Relembramos que este processo funciona completamente offline, não sendo permitido ao LLM executar RAG para enquadrar as respostas. É um modelo fixo e isolado, generalista, sem treino específico para Neonatologia.

As primeiras tentativas demonstraram alguma criatividade nas respostas, nem sempre assertivas ou sequer pragmáticas. Também se verificou que poderia ocorrer omissão de parâmetros e, como se esperaria, a mesma situação / caso nem sempre gerava a mesma conclusão / proposta, apesar dos dados serem sempre iguais. Assim, foi eliminada a estratégia "zero-shot".

A inclusão de um exemplo na prompt – "one-shot" – ou mesmo mais exemplos – "multiple-shot" – melhorou um pouco o desempenho, mas alguns problemas subsistiam: se algum dado faltava (o que pode acontecer nalguns textos do dataset), o LLM usava o do exemplo fornecido ou inventava um valor.

Assim, a estratégia foi reformulada várias vezes, sendo os melhores resultados produzidos com uma opção híbrida:

a) utilização dos mesmos parâmetros no LMStudio já aplicados na Tarefa 1 – Extração de dados: temperatura 0,15; top-K = 5 e top-p = 0,95
b) instruções rígidas na prompt
c) "one-shot" - inclusão de um exemplo para organização dos elementos na resposta
d) utilização de um JSON com todos os dados a incluir na conclusão

As instruções para a conclusão estão reproduzidas na Figura 61 – Prompt final para gerar uma conclusão e proposta de ajuste em linguagem natural

```python
def build_conclusion_prompt(final_json: dict, template_example: str) -> str:
    exemplo = (template_example or "").strip() or (
        "RN de 27 semanas e 3 dias com PN de 1200g (Z-Score -1,15). "
        "Atualmente em D14 de vida, com IPM de 29 semanas e 3 dias e Peso Atual de 1230g (Z-Score de -0,96). "
        "Tem um AHT de 145 ml/kg/dia, 120 Kcal/kg/dia, 3,2g/kg/dia de proteínas, 5g/kg/dia de lípidos, 10g/kg/dia de hidratos de carbono
        "com um PER de 2,9 e 4,2mg/kg/dia de ferro.\n"
        "Aportes adequados: AHT, calorias, PER, Ferro.\n"
        "Aportes a ajustar: Proteínas - aumentar 0,3g/kg/dia; hidratos de carbono - diminuir 0,2g/kg/dia."
    )
    regras = (
        "Objetivo: escrever a conclusão FINAL seguindo RIGIDAMENTE o template do exemplo.\n"
        "- NÃO acrescentar novas categorias, variáveis, ou interpretações.\n"
        "- NÃO omitir campos que existam no JSON fornecido.\n"
        "- Usar exatamente as unidades e a ordem do exemplo.\n"
        "- Se uma lista estiver vazia, escreve apenas o cabeçalho e deixa-a vazia.\n"
        "- Se algum valor em final_json['aportes'] for null (por exemplo calorias, proteínas, lípidos, hidratos de carbono, PER ou ferro
        "NÃO o escrevas na frase 'Tem um AHT de ...'.\n"
        "- Nesses casos, adiciona uma linha separada no fim com o texto exatamente:\n"
        "  'Dados em falta: ...'\n"
        "  onde '...' é a lista das variáveis em falta, por esta ordem fixa: calorias, proteínas, lípidos, hidratos, PER, ferro.\n"
        "- Se não houver valores em falta, NÃO escrevas a linha 'Dados em falta:'.\n"
        "- Não usar linguagem criativa; apenas factual com os valores do JSON."
    )
    dados = json.dumps(final_json, ensure_ascii=False)
    prompt = (
        regras + "\n\n"
        "TEMPLATE/EXEMPLO A SEGUIR:\n" + exemplo + "\n\n"
        "DADOS EM JSON PARA PREENCHER O TEMPLATE (usa-os TODOS, sem inventar):\n" + dados + "\n\n"
        "Produz APENAS o texto final, sem explicações."
    )
    return prompt
```

Figura 61 – Prompt final para gerar uma conclusão e proposta de ajuste em linguagem natural

O diagrama para a Tarefa 5 está reproduzido na Figura 62.

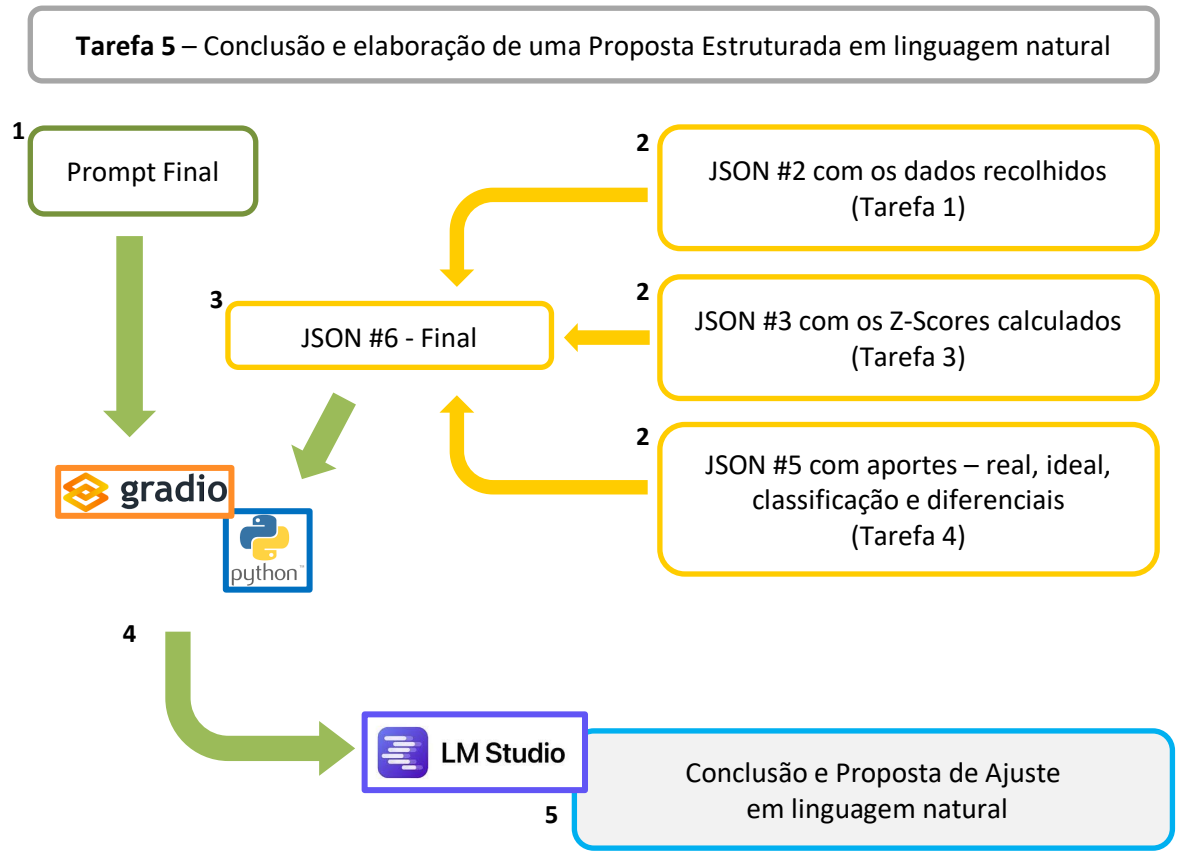


Figura 62 – Diagrama da Tarefa 5

A Tarefa 5 inicia-se com a conjugação de todos os dados produzidos ao longo do Processo Global. Através do Python é enviada uma prompt (1) contendo as instruções para a elaboração da Conclusão e Proposta, para além de um JSON #6 com a informação relevante dos anteriores (2-3). Toda esta informação é enviada ao LLM (4) através do LMStudio e o Python importa o texto gerado (5).

Na Figura 63 encontram-se alguns instantâneos da Tarefa 5.

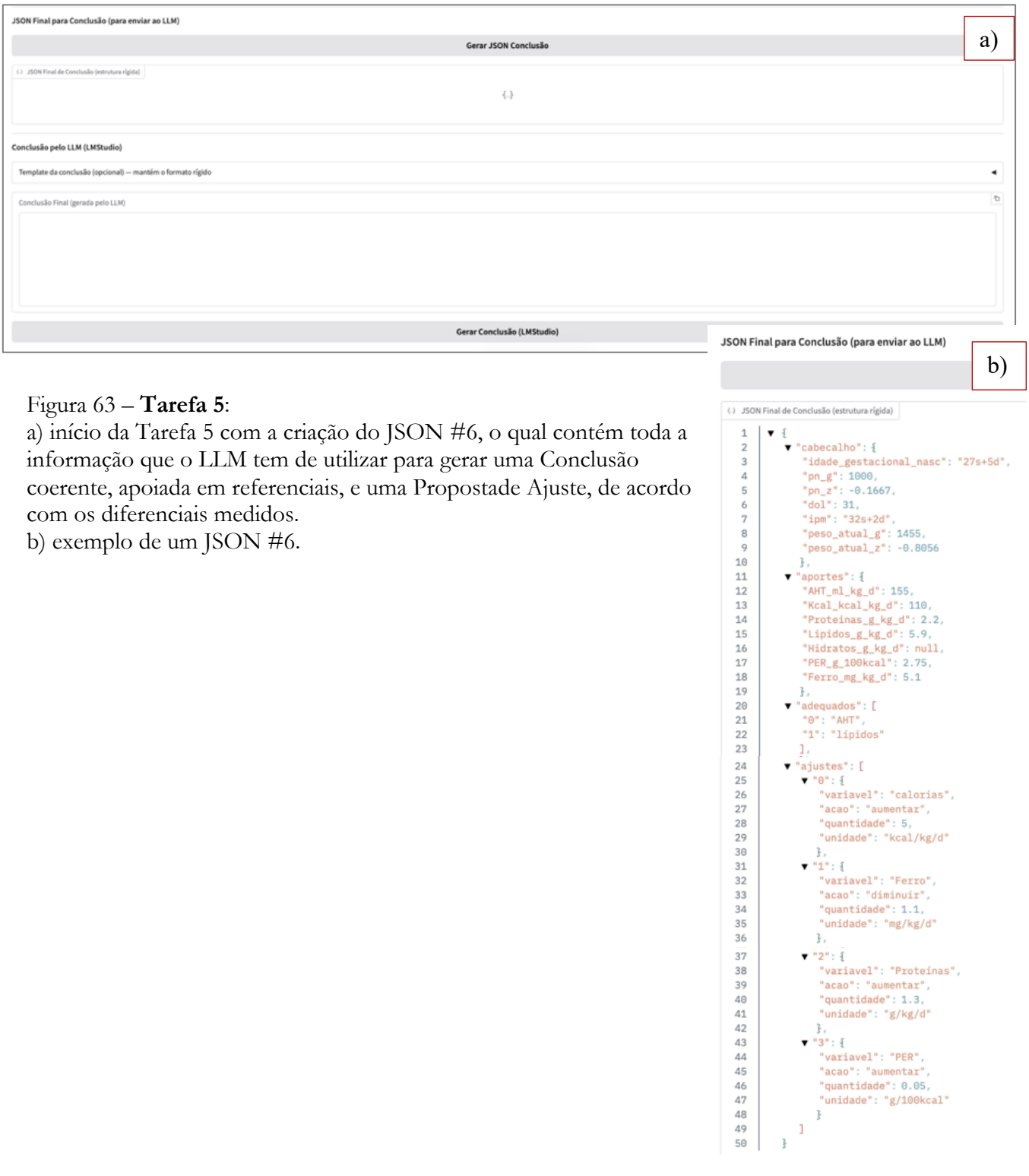


Figura 63 – **Tarefa 5**:
a) início da Tarefa 5 com a criação do JSON #6, o qual contém toda a informação que o LLM tem de utilizar para gerar uma Conclusão coerente, apoiada em referenciais, e uma Propostade Ajuste, de acordo com os diferenciais medidos.
b) exemplo de um JSON #6.

O texto final pode ser importado para introduzir no diário clínico, se a proposta é considerada válida e aceite pelo clínico:

RN de 27 semanas e 5 dias com PN de 1000g (Z-Score -0,1667). Atualmente em D31 de vida, com IPM de 32 semanas e 1 dias e Peso Atual de 1455g (Z-Score de -0,8056). Tem um AHT de 155 ml/kg/dia, 110 Kcal/kg/dia, 2,2g/kg/dia de proteínas, 5,9g/kg/dia de lípidos.

Aportes adequados: AHT, lípidos.

Aportes a ajustar: calorias - aumentar 5 kcal/kg/d; Ferro - diminuir 1,1 mg/kg/d; Proteínas - aumentar 1,3 g/kg/d; PER - aumentar 0,05 g/100kcal.

Dados em falta: hidratos.

Está assim terminado todo o processo.

8.3 Avaliação dos Resultados

No ponto anterior apresentámos a modelação do processo, em que se pretendia:

1) analisar um texto clínico de forma automatizada e autónoma;

2) extrair os itens necessários para estabelecer o estado atual dos aportes nutricionais;

3) comparar com as metas nutricionais individualizadas e ajustadas ao RN em questão;

4) classificar a adequação dos aportes reais e identificar desvios;

5) elaborar uma proposta de ajuste nutricional em linguagem natural.

Por fim, devemos analisar os resultados obtidos e avaliar o desempenho do processo:

- presença de todos os elementos nas várias tarefas – atribuído 1 ponto por item em cada tarefa; soma final.
- presença de todos os elementos na conclusão – atribuído 1 ponto por cada item que deve aparecer no texto final; soma final
- qualidade do texto – linguagem natural e concisa – atribuído 1 ponto para cada; criatividade na conclusão / não respeitar instruções / elementos extra / tentativas de raciocínio – 1 ponto negativo por cada; soma final

As pontuações estão apresentadas na Tabela 11 – Resultados obtidos na amostra com 30 textos clínicos provenientes do dataset.

Tabela 11 – Resultados obtidos na amostra com 30 textos clínicos provenientes do dataset

Textos	Tarefa 1	Tarefa 2	Tarefa 3	Tarefa 4	Tarefa 5	Qualidade	Total
	Extração elementos 0 - 15	Atribuição Z-Score 0 - 4	Standard Nutricional 0 - 7	Comparação e Diferenciais 0 - 7	Elementos obrigatórios 0 - 14	Linguagem e Criatividade	
RN 31.txt	15	4	7	7	14	2	49
RN 105.txt	15	4	7	7	14	2	49
RN 175.txt	15	4	7	7	14	2	49
RN 61.txt	15	4	7	7	14	2	49
RN 169.txt	15	4	7	7	14	2	49
RN 108.txt	15	4	7	7	14	2	49
RN 01.txt	15	4	7	7	14	2	49
RN 74.txt	15	4	7	7	14	2	49
RN 83.txt	15	4	7	7	14	2	49
RN 46.txt	10	1	3	3	7	2	26
RN 51.txt	15	4	7	7	14	2	49
RN 08.txt	15	4	7	7	14	2	49
RN 102.txt	15	4	7	7	14	2	49
RN 166.txt	15	4	7	7	14	2	49
RN 23.txt	15	4	7	7	14	2	49
RN 03.txt	15	4	7	7	14	2	49
RN 44.txt	10	1	3	3	7	2	26
RN 96.txt	15	4	7	7	14	2	49
RN 170.txt	15	4	7	7	14	2	49
RN 57.txt	15	4	7	7	14	2	49
RN 32.txt	15	4	7	7	14	2	49
RN 88.txt	15	4	7	7	14	2	49
RN 79.txt	15	4	7	7	14	2	49
RN 153.txt	15	4	7	7	14	2	49
RN 41.txt	10	1	3	3	7	2	26
RN 92.txt	15	4	7	7	14	2	49
RN 10.txt	15	4	7	7	14	2	49
RN 37.txt	10	1	3	3	7	2	26
RN 28.txt	15	4	7	7	14	2	49
RN 93.txt	15	4	7	7	14	2	49

A Avaliação dos Resultados integra-se na lógica já referida no CRISP-MED-DM e aqui utilizada para estruturar o processo.

Os resultados apresentados na Tabela 11 – Resultados obtidos na amostra com 30 textos clínicos provenientes do dataset – são favoráveis e demonstram um desempenho aceitável do modelo desenvolvido.

Em quatro dos textos da amostra (assinalados a cor) foram obtidos resultados totais inferiores aos restantes. Esta diferença não se deve ao funcionamento do Processo, mas sim à ausência das variáveis nos textos analisados. Ou seja, algumas variáveis não são extraídas porque não existem; logo todas as tarefas do Processo que as manipulam contam com menos itens pontuáveis.

Podemos também referir que os resultados são, pelo menos, não inferiores ao gold standard, que, à falta de melhor, é o modelo tradicional manual humano.

Como inovação, podemos evidenciar o carácter determinístico (e auditável) do processamento via Python, dos dados extraídos o que pode diminuir marcadamente o erro humano. Ainda assim, não podemos confiar cegamente na conclusão

apresentada, porque esta depende de duas passagens por um LLM (por definição, probabilístico e não determinístico).

A primeira passagem pelo LLM é a que recolhe os dados do texto clínico, e apesar dos bons resultados do modelo escolhido após um processo exaustivo de avaliação, este continua a ser probabilístico. A segunda passagem pelo LLM é a fase final: apesar das precauções para minimizar a criatividade e manter um texto coerente, completo e protocolado, baseado num JSON final com a informação relevante, haverá sempre a possibilidade de alguma imprecisão.

Daí a necessidade de validação humana e decisão de aceitar ou não a Conclusão e Proposta.

Todas as cinco Tarefas do processo podem ser melhoradas, com especial ênfase nas duas que dependem diretamente do LLM.

Por último, todo o processo pode ser amplamente melhorado, condensando as cinco tarefas de modo sequencial e automatizado, sem necessidade de intervenção para avançar entre Tarefas.

9 DISCUSSÃO

9.1 Contribuições

Concluído o projeto, pode-se afirmar que o objetivo proposto foi cumprido na íntegra, ou mesmo superado, atendendo aos contributos para o estado da arte listados abaixo.

De uma maneira geral, concluímos favoravelmente todas as etapas do Projeto.

A aplicação da IA no contexto de uma UCIN é possível e pode trazer vantagens, quer na eliminação de erro humano, quer como "assistente" na prática clínica - automatizando (e autonomizando) ações – quer como facilitador de tarefas repetitivas diárias.

Não existem trabalhos publicados explorando a utilização de um LLM simultaneamente como aplicação de NER e como assistente clínico; ou seja, tal como já referimos, que possa extrair dados e autonomamente propor decisões clínicas, sem recorrer a modelos online *state of the art.*

Esta solução, executável em sistemas razoavelmente comuns, permite solucionar muitos dos problemas ligados a segurança de dados clínicos. Neste trabalho todos os dados permanecem na rede hospitalar e, como o modelo está fixo *(frozen),* não há lugar a memória dos dados.

Mais especificamente, provou-se que uma aplicação utilizando um LLM pode extrair elementos clínicos e propor um ajuste na nutrição, sem intervenção humana; apenas bastando aprovar a proposta.

Assim, esta abordagem é pioneira nas vertentes da segurança dos dados, na utilização de um LLM *onsite*, na aplicação em Neonatologia e na exploração da vertente nutricional em prematuros.

A metodologia utilizada, desde a organização baseada em CRISP-MED-DM, a necessidade de criar um dataset sintético, a validação do mesmo, todo o processo de teste dos modelos LLM para escolher o mais ajustado até à organização de todo o Processo são mais valias para esta temática.

Paralelamente, foi já publicado um artigo sobre este Projeto, ainda que abordasse apenas a fase inicial como *proof of concept* [104]:

Rui Castelo, Mateus Mendes. Adapting Large Language Models to Support Nutritional Decision-Making in Neonatal Intensive Care. In Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

Seguidamente, será elaborado um segundo artigo com o processo completo.

9.2 Resposta às Questões da Investigação

No âmbito desta Tese, o papel do LLM como assistente de atividade clínica foi testado num cenário específico:

"Qual a capacidade de um LLM para elaborar uma proposta de ajuste nutricional com base no diário clínico do RN em UCIN?"

E com o Objetivo da Tese:

"Testar a capacidade de um LLM para elaborar uma proposta de ajuste nutricional individualizada, partindo do texto do diário clínico, num processo limitado pelo hardware habitual na ULS e excluindo qualquer processamento online."

Assim, podemos concluir que:

1) O LLM demonstrou ser capaz de funcionar como assistente de atividade clínica ao conseguir elaborar uma proposta de ajuste nutricional com base na informação que recolheu do texto do diário clínico.
2) Neste Projeto, conseguimos demonstrar que se pode executar um processo de extração de elementos clínicos e sua utilização para criar uma proposta de ajuste terapêutico / nutricional, baseado em LLM, e estando limitado a equipamento informático genérico.

Assim, está encontrada a resposta à Questão de Investigação e cumprido o Objetivo da Tese.

9.3 Implementação futura em UCIN

A fase final do processo, de acordo com a metodologia CRISP-MED-DM seguida neste Projeto, compreenderia a Implementação em ambiente real – a UCIN, a explorar numa eventual extensão e aprofundamento em novo Projeto, que se encontra fora do âmbito desta Tese.

Como é de esperar, esta fase dependerá de autorização de várias entidades, desde a Direção de Departamento, a Comissão de Ética e o Encarregado de Proteção de Dados (DPO) da ULS.

A título especulativo, podemos extrapolar e tecer algumas considerações.

Na fase de Implementação, toda a arquitetura do processo poderá ter de ser revista, especialmente na simplificação da intervenção do operador (diminuindo o número de cliques e carregamento de arquivos necessários - estes seriam sempre os mesmos). Idealmente, deveria ser apenas necessário introduzir o texto clínico e obter uma conclusão com a proposta de ajuste. Todo o processamento teria de estar oculto, obviando-se a necessidade de gerir os vários programas utilizados.

Também é importante avaliar a opinião dos clínicos sobre o processo em geral.

Por fim, é necessário monitorizar todos os resultados, mantendo um acompanhamento de todo o processo, garantindo a fiabilidade dos resultados e a atualização de protocolos e standards.

Deverá também ser alvo de um relatório periódico.

Em resumo, a fase final deve ser entendida também como um Ciclo PDCA clássico.

9.4 Perspetivas Futuras

Após a conclusão deste Projeto, ficam ainda muitas possibilidades para explorar.

Demonstrou-se que um LLM *offline*, confinado à rede da Instituição, pode extrair dados de textos clínicos não estruturados e utilizar esses dados para tecer conclusões.

Esta abordagem pode ser utilizada para outras áreas de cuidados intensivos neonatais, como o crescimento, a ventilação, parâmetros hematológicos, etc. Pode ainda ser aplicada a áreas como a gestão e eventualmente a bases de dados.

10 CONCLUSÃO

Após este caminho, voltamos à questão inicial:

"Qual a capacidade de um LLM para elaborar uma proposta de ajuste nutricional com base em diários clínicos?"

Para responder a esta questão, foi necessário:

1) Definir um processo global que mimetizasse o pensamento humano para resolver este problema.
2) Decompor esse processo em tarefas menores, cada uma com especificidades diferentes, utilizando várias aplicações:
    a. análise e extração de variáveis de um texto clínico, semi-estruturado, através de um LLM. Definição do melhor método para utilizar o LLM, ajustando parâmetros, prompt, etc. Possibilidade de evolução futura das variáveis a recolher.
    b. comparação dessas variáveis com standards e composição de variáveis derivadas. Possibilidade de atualização futura dos standards.
    c. utilização de todos os dados (extraídos, processados e compostos) pelo LLM para gerar uma conclusão em linguagem natural.
3) Como componentes adicionais, mas incontornáveis, deste projeto, destacamos a descrição de todo o projeto segundo a metodologia CRISP-MED-DM; a criação de um dataset sintético para teste do modelo; a criação de um fluxo de extração de variáveis e, por fim, um fluxo para avaliação de vários LLM para selecionar o mais adequado à tarefa.
4) Contribuímos para o estado da arte, com um artigo já publicado (abordando o *proof of concept)* [104]:

Rui Castelo, Mateus Mendes. Adapting Large Language Models to Support Nutritional Decision-Making in Neonatal Intensive Care. In Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

5) Seguidamente, será elaborado um segundo artigo com o processo completo.

A principal conclusão é a de que a IA, neste caso através de um LLM pode, de facto, ajudar na avaliação e ajuste de aportes nutricionais em RN numa UCIN. No processo proposto, existe a possibilidade de evoluir na lista de variáveis a recolher e nos standards para comparação.

Este papel da IA apresenta como principal vantagem a automatização e rapidez na análise de aportes nutricionais, e ainda permite obviar algumas limitações humanas, como o cansaço, a divisão da atenção, as interrupções, a generalização, etc.

Mas devemos também referir que os resultados favoráveis obtidos com o processo apresentado nesta Tese não devem ser, cegamente, generalizados. Ou seja, o papel dos LLM com NER deve ser sempre testado no contexto específico, bem como outras funções adicionais.

Ainda assim, a sua utilização, controlada, como NER permite abrir mais uma porta na área da análise de texto clínico, com todas as possibilidades futuras. Após recolher dados de forma assertiva, poderemos considerar múltiplas aplicações com os mesmos.

11 REFERÊNCIAS BIBLIOGRÁFICAS

[1] D. Rallis, M. Baltogianni, K. Kapetaniou, and V. Giapros, “Current Applications of Artificial Intelligence in the Neonatal Intensive Care Unit,” Jun. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/biomedinformatics4020067.

[2] McCarthy J, Minsky ML, Rochester N, and Shannon CE, “A proposal for the Dartmouth Summer Research Project on Artificial Intelligence,” Aug. 1955.

[3] C. Li, R. Zhou, G. Chen, X. Hao, and T. Zhu, “Knowledge mapping and research hotspots of artificial intelligence on ICU and Anesthesia: from a global bibliometric perspective,” *Anesthesiology and Perioperative Science*, vol. 1, no. 4, Oct. 2023, doi: 10.1007/s44254-023-00031-5.

[4] A. Bottrighi and M. Pennisi, “Exploring the State of Machine Learning and Deep Learning in Medicine: A Survey of the Italian Research Community,” Sep. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/info14090513.

[5] K. Beam, P. Sharma, P. Levy, and A. L. Beam, “Artificial intelligence in the neonatal intensive care unit: the time is now,” *Journal of Perinatology*, vol. 44, no. 1, pp. 131–135, Jan. 2024, doi: 10.1038/s41372-023-01719-z.

[6] W. A. Knaus, E. A. Draper, D. P. Wagner, and J. E. Zimmerman, “APACHE II: a severity of disease classification system.,” *Crit Care Med*, vol. 13, no. 10, pp. 818–29, Oct. 1985.

[7] B. A. Sullivan *et al.*, “Transforming neonatal care with artificial intelligence: challenges, ethical consideration, and opportunities,” Jan. 01, 2024, *Springer Nature*. doi: 10.1038/s41372-023-01848-5.

[8] J. Gotman A’, D. Flanagan, J. Zhang, and B. Rosenblatt, “Automatic seizure detection in the newborn: methods and initial evaluation,” 1997.

[9] S. A. Coggins and K. Glaser, “Updates in Late-Onset Sepsis: Risk Assessment, Therapy, and Outcomes,” Nov. 01, 2022, *NLM (Medline)*. doi: 10.1542/neo.23-10-e738.

[10] D. D. Flannery, E. M. Edwards, S. A. Coggins, J. D. Horbar, and K. M. Puopolo, “Late-Onset Sepsis Among Very Preterm Infants,” *Pediatrics*, vol. 150, no. 6, Dec. 2022, doi: 10.1542/PEDS.2022-058813.

[11] M.-H. Tsai *et al.*, “Incidence, Clinical Characteristics and Risk Factors for Adverse Outcome in Neonates With Late-onset Sepsis,” *Pediatr Infect Dis J*, vol. 33, no. 1, 2014, [Online]. Available: https://journals.lww.com/pidj/fulltext/2014/01000/incidence,_clinical_characteristics_and_risk.8.aspx

[12] B. Meskó and M. Görög, “A short guide for medical professionals in the era of artificial intelligence,” Dec. 01, 2020, *Nature Research*. doi: 10.1038/s41746-020-00333-z.

[13] E. Keles and U. Bagci, “The past, current, and future of neonatal intensive care units with artificial intelligence: a systematic review,” Dec. 01, 2023, *Nature Research*. doi: 10.1038/s41746-023-00941-5.

[14] W. Song, S. Y. Jung, H. Baek, C. W. Choi, Y. H. Jung, and S. Yoo, “A predictive model based on machine learning for the early detection of late-onset neonatal sepsis: Development and observational study,” *JMIR Med Inform*, vol. 8, no. 7, Jul. 2020, doi: 10.2196/15965.

[15] A. Kallonen, M. Juutinen, A. Värri, G. Carrault, P. Pladys, and A. Beuchée, "Early detection of late-onset neonatal sepsis from noninvasive biosignals using deep learning: A multicenter prospective development and validation study," *Int J Med Inform*, vol. 184, Apr. 2024, doi: 10.1016/j.ijmedinf.2024.105366.

[16] S. Mani *et al.*, "Medical decision support using machine learning for early detection of late-onset neonatal sepsis," *Journal of the American Medical Informatics Association*, vol. 21, no. 2, pp. 326–336, Mar. 2014, doi: 10.1136/amiajnl-2013-001854.

[17] M. Yang *et al.*, "Continuous prediction and clinical alarm management of late-onset sepsis in preterm infants using vital signs from a patient monitor," *Comput Methods Programs Biomed*, vol. 255, Oct. 2024, doi: 10.1016/j.cmpb.2024.108335.

[18] M. Meeus *et al.*, "Clinical Decision Support for Improved Neonatal Care: The Development of a Machine Learning Model for the Prediction of Late-onset Sepsis and Necrotizing Enterocolitis," *Journal of Pediatrics*, vol. 266, Mar. 2024, doi: 10.1016/j.jpeds.2023.113869.

[19] A. J. Masino *et al.*, "Machine learning models for early sepsis recognition in the neonatal intensive care unit using readily available electronic health record data," *PLoS One*, vol. 14, no. 2, Feb. 2019, doi: 10.1371/journal.pone.0212665.

[20] K. D. Fairchild *et al.*, "Vital signs and their cross-correlation in sepsis and NEC: A study of 1,065 very-low-birth-weight infants in two NICUs," *Pediatr Res*, vol. 81, no. 2, pp. 315–321, Feb. 2017, doi: 10.1038/pr.2016.215.

[21] D. Van Laere *et al.*, "Machine Learning to Support Hemodynamic Intervention in the Neonatal Intensive Care Unit," Sep. 01, 2020, *W.B. Saunders*. doi: 10.1016/j.clp.2020.05.002.

[22] E. Korot *et al.*, "Predicting sex from retinal fundus photographs using automated deep learning," *Sci Rep*, vol. 11, no. 1, Dec. 2021, doi: 10.1038/s41598-021-89743-x.

[23] S. Shao, T. Wang, C. Song, X. Chen, E. Cui, and H. Zhao, "Obstructive sleep apnea recognition based on multi-bands spectral entropy analysis of short-time heart rate variability," *Entropy*, vol. 21, no. 8, 2019, doi: 10.3390/e21080812.

[24] F.-T.-Z. Khanam *et al.*, "Automatic Vital Signs Monitoring of Infants in A Neonatal Intensive Care Unit Based on Neural Networks," *J. Imaging*, vol. 2021, p. 122, 2021, doi: 10.3390/jimaging.

[25] I. Lorato *et al.*, "Towards continuous camera-based respiration monitoring in infants," *Sensors*, vol. 21, no. 7, Apr. 2021, doi: 10.3390/s21072268.

[26] M. Villarroel *et al.*, "Non-contact physiological monitoring of preterm infants in the Neonatal Intensive Care Unit," *NPJ Digit Med*, vol. 2, no. 1, Dec. 2019, doi: 10.1038/s41746-019-0199-5.

[27] I. Lorato *et al.*, "Multi-camera infrared thermography for infant respiration monitoring," *Biomed Opt Express*, vol. 11, no. 9, p. 4848, Sep. 2020, doi: 10.1364/boe.397188.

[28] V. Karlsson, Y. T. Blomqvist, and J. Ågren, "Nursing care of infants born extremely preterm," Jun. 01, 2022, *W.B. Saunders Ltd*. doi: 10.1016/j.siny.2022.101369.

[29] T. Zhu, Y. Yang, J. Tang, and T. Xiong, "Machine learning for predicting intraventricular hemorrhage in preterm infants," *J Evid Based Med*, vol. 17, no. 1, pp. 7–9, Mar. 2024, doi: https://doi.org/10.1111/jebm.12561.

[30] A. Ortega-Leon, J. Pizarro, I. Benavente Fernández, S. P. Lubián López, and L. C. Gontard, "Automatic neonatal cranial ultrasound segmentation using deep learning: A review," 2022. [Online]. Available: http://ceur-ws.org

[31] N. Fatima *et al.*, “Deep learning approaches for automated classification of neonatal lung ultrasound with assessment of human-to-AI interrater agreement,” *Comput Biol Med*, vol. 183, Dec. 2024, doi: 10.1016/j.compbiomed.2024.109315.

[32] U. Khan *et al.*, “TranSLUCEnT: Transferred Sequential Lung Ultrasound Characteristic Encodings-based Transformer for Lung Ultrasound Pattern Classification in Premature Neonates,” in *IEEE Ultrasonics, Ferroelectrics, and Frequency Control Joint Symposium, UFFC-JS 2024 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/UFFC-JS60046.2024.10793539.

[33] L. Pezza *et al.*, “Meta-Analysis of Lung Ultrasound Scores for Early Prediction of Bronchopulmonary Dysplasia,” Apr. 01, 2022, *American Thoracic Society*. doi: 10.1513/AnnalsATS.202107-822OC.

[34] S. Aujla, A. Mohammed, N. Khan, and K. Umapathy, “Multi-Level Classification of Lung Pathologies in Neonates using Recurrence Features,” in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 1531–1535. doi: 10.1109/EMBC48229.2022.9871011.

[35] R. Bassiouny, A. Mohamed, K. Umapathy, and N. Khan, “An Interpretable Object Detection-Based Model for the Diagnosis of Neonatal Lung Diseases Using Ultrasound Images,” in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 3029–3034. doi: 10.1109/EMBC46164.2021.9630169.

[36] J. Erno, T. Gomes, C. Baltimore, J. P. Lineberger, D. H. Smith, and G. H. Baker, “Automated Identification of Patent Ductus Arteriosus Using a Computer Vision Model,” *Journal of Ultrasound in Medicine*, vol. 42, no. 12, pp. 2707–2713, Dec. 2023, doi: https://doi.org/10.1002/jum.16305.

[37] A. Gearhart, S. Goto, R. C. Deo, and A. J. Powell, “An Automated View Classification Model for Pediatric Echocardiography Using Artificial Intelligence,” *Journal of the American Society of Echocardiography*, vol. 35, no. 12, pp. 1238–1246, Dec. 2022, doi: 10.1016/j.echo.2022.08.009.

[38] R. D. E. Henderson, X. Yi, S. J. Adams, and P. Babyn, “Automatic Detection and Classification of Multiple Catheters in Neonatal Radiographs with Deep Learning,” *J Digit Imaging*, vol. 34, no. 4, pp. 888–897, Aug. 2021, doi: 10.1007/s10278-021-00473-y.

[39] L. He *et al.*, “Deep Multimodal Learning From MRI and Clinical Data for Early Prediction of Neurodevelopmental Deficits in Very Preterm Infants,” *Front Neurosci*, vol. 15, Oct. 2021, doi: 10.3389/fnins.2021.753033.

[40] A. Natarajan, G. Lam, J. Liu, A. L. Beam, K. S. Beam, and J. C. Levin, “Prediction of extubation failure among low birthweight neonates using machine learning,” *Journal of Perinatology*, vol. 43, no. 2, pp. 209–214, Feb. 2023, doi: 10.1038/s41372-022-01591-3.

[41] A. Mikhno and C. M. Ennett, “Prediction of extubation failure for neonates with respiratory distress syndrome using the MIMIC-II clinical database,” in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2012, pp. 5094–5097. doi: 10.1109/EMBC.2012.6347139.

[42] M. Fevereiro-Martins, C. Marques-Neves, H. Guimarães, and M. Bicho, “Retinopathy of prematurity: A review of pathophysiology and signaling pathways,” Mar. 01, 2023, *Elsevier Inc.* doi: 10.1016/j.survophthal.2022.11.007.

[43] L. F. Nakayama *et al.*, "Fairness and generalisability in deep learning of retinopathy of prematurity screening algorithms: A literature review," *BMJ Open Ophthalmol*, vol. 8, no. 1, Aug. 2023, doi: 10.1136/bmjophth-2022-001216.

[44] Z. Luo, X. Ding, N. Hou, and J. Wan, "A Deep-Learning-Based Collaborative Edge–Cloud Telemedicine System for Retinopathy of Prematurity," *Sensors*, vol. 23, no. 1, Jan. 2023, doi: 10.3390/s23010276.

[45] J. P. Campbell *et al.*, "Evaluation of a Deep Learning–Derived Quantitative Retinopathy of Prematurity Severity Scale," *Ophthalmology*, vol. 128, no. 7, pp. 1070–1076, Jul. 2021, doi: 10.1016/j.ophtha.2020.10.025.

[46] Q. Wu *et al.*, "Development and Validation of a Deep Learning Model to Predict the Occurrence and Severity of Retinopathy of Prematurity," *JAMA Netw Open*, vol. 5, no. 6, p. E2217447, Jun. 2022, doi: 10.1001/jamanetworkopen.2022.17447.

[47] F. Zhao, C. Zhang, K. M. Dudding, A. N. Sanders, P. Lewis-Chumley, and L. Kathryn, "Neonatal Pain Detection from Facial Expressions Using Deep Learning," Feb. 29, 2024. doi: 10.21203/rs.3.rs-3979706/v1.

[48] L. Fontes Buzuti, T. M. Heideirich, M. C. M. Barros, R. Guinsburg, and C. E. Thomaz, "Neonatal Pain Assessment From Facial Expression Using Deep Neural Networks."

[49] V. Giordano *et al.*, "Comparative analysis of artificial intelligence and expert assessments in detecting neonatal procedural pain," *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-71278-6.

[50] *2019 IEEE Healthcare Innovations and Point of Care Technologies, (HI-POCT)*. IEEE, 2019.

[51] A. Erdogan Yildirim and M. Canayaz, "Machine learning-based prediction of length of stay (LoS) in the neonatal intensive care unit using ensemble methods," *Neural Comput Appl*, vol. 36, no. 23, pp. 14433–14448, Aug. 2024, doi: 10.1007/s00521-024-09831-7.

[52] N. D. Embleton *et al.*, "Enteral Nutrition in Preterm Infants (2022): A Position Paper from the ESPGHAN Committee on Nutrition and Invited Experts," *J Pediatr Gastroenterol Nutr*, vol. 76, no. 2, pp. 248–268, Feb. 2023, doi: 10.1097/MPG.0000000000003642.

[53] M. Meiliana *et al.*, "Nutrition guidelines for preterm infants: A systematic review," Jan. 01, 2024, *John Wiley and Sons Inc*. doi: 10.1002/jpen.2568.

[54] L. Pereira-Da-silva *et al.*, "Guidelines for enteral nutrition in infants born preterm: 2023 update by the Portuguese Neonatal Society. Part II. Enteral feeding in specific clinical conditions and feeding after discharge," *Portuguese Journal of Pediatrics*, vol. 54, no. 4, pp. 264–270, Oct. 2023, doi: 10.24875/PJP.23000005.

[55] L. Pereira-Da-silva *et al.*, "Guidelines for enteral nutrition in infants born preterm: 2023 update by the Portuguese Neonatal Society. Part I. Nutrient requirements and enteral feeding approach during the hospital stay," *Portuguese Journal of Pediatrics*, vol. 54, no. 4, pp. 253–263, Oct. 2023, doi: 10.24875/PJP.23000004.

[56] V. Ntinopoulos *et al.*, "Large language models for data extraction from unstructured and semi-structured electronic health records: A multiple model performance evaluation," *BMJ Health Care Inform*, vol. 32, no. 1, Jan. 2025, doi: 10.1136/bmjhci-2024-101139.

[57] Y. Wang *et al.*, "Clinical information extraction applications: A literature review," Jan. 01, 2018, *Academic Press Inc.* doi: 10.1016/j.jbi.2017.11.011.

[58] A. R. Aronson and F. M. Lang, “An overview of MetaMap: Historical perspective and recent advances,” *Journal of the American Medical Informatics Association*, vol. 17, no. 3, pp. 229–236, May 2010, doi: 10.1136/jamia.2009.002733.
[59] G. K. Savova *et al.*, “Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications,” *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 507–513, Sep. 2010, doi: 10.1136/jamia.2009.001560.
[60] A. Aronson, “Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program,” *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, vol. 2001, pp. 17–21, Feb. 2001.
[61] B. Settles, “Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets.” [Online]. Available: http://us.expasy.org/sprot/
[62] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” [Online]. Available: https://github.com/tensorflow/tensor2tensor
[63] E. Alsentzer *et al.*, “Publicly Available Clinical BERT Embeddings,” 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/
[64] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” [Online]. Available: https://github.com/tensorflow/tensor2tensor
[65] S. Abedian *et al.*, “Automated Extraction of Tumor Staging and Diagnosis Information From Surgical Pathology Reports,” *JCO Clin Cancer Inform*, vol. 5, pp. 1054–1061, 2021, doi: 10.1200/CCI.21.
[66] H. Wang *et al.*, “Using natural language processing in emergency medicine health service research: A systematic review and meta-analysis,” Jul. 01, 2024, *John Wiley and Sons Inc*. doi: 10.1111/acem.14937.
[67] H. Zhou *et al.*, “A Survey of Large Language Models in Medicine: Progress, Application, and Challenge.” [Online]. Available: https://github.com/AI-in-Health/MedLLMsPracticalGuide.
[68] H. Touvron *et al.*, “LLaMA: Open and Efficient Foundation Language Models,” Feb. 2023, [Online]. Available: http://arxiv.org/abs/2302.13971
[69] Z. Yang *et al.*, “Large Language Model–Based Critical Care Big Data Deployment and Extraction: Descriptive Analysis,” *JMIR Med Inform*, vol. 13, 2025, doi: 10.2196/63216.
[70] H. Zhou *et al.*, “A Survey of Large Language Models in Medicine: Progress, Application, and Challenge.” [Online]. Available: https://github.com/AI-in-Health/MedLLMsPracticalGuide.
[71] T. Young, D. Hazarika, S. Poria, and E. Cambria, “Recent trends in deep learning based natural language processing [Review Article],” Aug. 01, 2018, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/MCI.2018.2840738.
[72] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, “Large language models in medicine,” Aug. 01, 2023, *Nature Research*. doi: 10.1038/s41591-023-02448-8.
[73] A. Vaswani *et al.*, “Attention Is All You Need,” Jun. 2017, [Online]. Available: http://arxiv.org/abs/1706.03762
[74] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” Sep. 2013, [Online]. Available: http://arxiv.org/abs/1301.3781
[75] S. Lynn, “An Introduction to Word Embeddings for Text Analysis,” Oct. 2018.

[76] K. Choromanski *et al.*, “Rethinking Attention with Performers,” Nov. 2022, [Online]. Available: http://arxiv.org/abs/2009.14794
[77] J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer Normalization,” Jul. 2016, [Online]. Available: http://arxiv.org/abs/1607.06450
[78] C. Raffel *et al.*, “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,” Sep. 2023, [Online]. Available: http://arxiv.org/abs/1910.10683
[79] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” [Online]. Available: https://github.com/tensorflow/tensor2tensor
[80] E. Alsentzer *et al.*, “Publicly Available Clinical BERT Embeddings,” 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/
[81] L. Boonstra *et al.*, *Prompt Engineering*. Google Cloud, 2024.
[82] S. C. Lee, D. G. Lee, and Y. S. Seo, “Determining the best feature combination through text and probabilistic feature analysis for GPT-2-based mobile app review detection,” *Applied Intelligence*, vol. 54, no. 2, pp. 1219–1246, Jan. 2024, doi: 10.1007/s10489-023-05201-3.
[83] J. Long, “Large Language Model Guided Tree-of-Thought,” May 2023, [Online]. Available: http://arxiv.org/abs/2305.08291
[84] R. Sapkota, K. I. Roumeliotis, and M. Karkee, “AI Agents vs. Agentic AI: A Conceptual Taxonomy, Applications and Challenges,” May 2025, doi: 10.1016/j.inffus.2025.103599.
[85] M. Cheng *et al.*, “A Survey on Knowledge-Oriented Retrieval-Augmented Generation,” Mar. 2025, [Online]. Available: http://arxiv.org/abs/2503.10677
[86] L. D. X. L. S. Z. X. S. S. W. J. L. R. H. T. Z. F. W. G. W. Shengyu Zhang, “Instruction Tuning for Large Language Models: A Survey,” Aug. 2025, doi: https://doi.org/10.48550/arXiv.2308.10792.
[87] Gang Liu, Zhaolin Chen, Genrong He, Jinlong He, Pengfei Li, and Shenjun Zhong, “PeFoMed: Parameter Efficient Fine-tuning of Multimodal Large Language Models for Medical Imaging,” Apr. 2024, pp. 1–12.
[88] L. Wang *et al.*, “Parameter-efficient fine-tuning in large language models: a survey of methodologies,” *Artif Intell Rev*, vol. 58, no. 8, Aug. 2025, doi: 10.1007/s10462-025-11236-4.
[89] IBM, “What are AI hallucinations?”
[90] A. T. Kalai, O. Ofir, N. Openai, S. S. Vempala, G. Tech, and E. Z. Openai, “Why Language Models Hallucinate,” 2025. [Online]. Available: https://meta.ai
[91] T. Suenghataiphorn, N. Tribuddharat, P. Danpanichkul, and N. Kulthamrongsri, “Bias in Large Language Models Across Clinical Applications: A Systematic Review.” doi: https://doi.org/10.48550/arXiv.2504.02917.
[92] M. Omar *et al.*, “Evaluating and addressing demographic disparities in medical large language models: a systematic review,” Dec. 01, 2025, *BioMed Central Ltd*. doi: 10.1186/s12939-025-02419-0.
[93] A. Mahajan, Z. Obermeyer, R. Daneshjou, J. Lester, and D. Powell, “Cognitive bias in clinical large language models,” Dec. 01, 2025, *Nature Research*. doi: 10.1038/s41746-025-01790-0.
[94] C. Singh, J. P. Inala, M. Galley, R. Caruana, and J. Gao, “Rethinking Interpretability in the Era of Large Language Models,” Jan. 2024, [Online]. Available: http://arxiv.org/abs/2402.01761

[95] P. Shojaee, I. Mirzadeh, and K. Alizadeh Maxwell Horton Samy Bengio Mehrdad Farajtabar Apple, “The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity.”
[96] D. Plunkett, A. Morris, K. Reddy, and J. Morales, “Self-Interpretability: LLMs Can Describe Complex Internal Processes that Drive Their Decisions, and Improve with Training,” May 2025, [Online]. Available: http://arxiv.org/abs/2505.17120
[97] X. Wang, M. Salmani, P. Omidi, X. Ren, M. Rezagholizadeh, and A. Eshaghi, “Beyond the Limits: A Survey of Techniques to Extend the Context Length in Large Language Models,” May 2024, [Online]. Available: http://arxiv.org/abs/2402.02244
[98] J. Jonnagaddala and Z. S. Y. Wong, “Privacy preserving strategies for electronic health records in the era of large language models,” Dec. 01, 2025, *Nature Research*. doi: 10.1038/s41746-025-01429-0.
[99] M. Williams, W. Karim, J. Gelman, and M. Raza, “Ethical data acquisition for LLMs and AI algorithms in healthcare,” *NPJ Digit Med*, vol. 7, no. 1, Dec. 2024, doi: 10.1038/s41746-024-01399-9.
[100] R. Wirth and J. Hipp, “CRISP-DM: Towards a Standard Process Model for Data Mining.”
[101] O. Niaksu and O. Niakšu, “CRISP Data Mining Methodology Extension for Medical Domain,” 2015. [Online]. Available: https://www.researchgate.net/publication/277775478
[102] D. M. W. Powers and Ailab, “EVALUATION: FROM PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION.”
[103] R. J. Hyndman and A. B. Koehler, “Another look at measures of forecast accuracy,” *Int J Forecast*, vol. 22, no. 4, pp. 679–688, Oct. 2006, doi: 10.1016/j.ijforecast.2006.03.001.
[104] R. Castelo and M. Mendes, “Adapting Large Language Models to Support Nutritional Decision-Making in Neonatal Intensive Care.” [Online]. Available: https://pamdas.rcm2.pt/event/1/

Anexos

Anexo A – artigo científico realizado na fase exploratória do processo

Proceedings of PAMDAS 2025 - International Conference on Physical Asset Management and Data Science.
Coimbra, Portugal, 17/18 July 2025. https://pamdas.rcm2.pt/event/1/

Adapting Large Language Models to Support Nutritional Decision-Making in Neonatal Intensive Care

Rui Castelo[1,2,*], Mateus Mendes[1,3]
[1] Polytechnic University of Coimbra, Coimbra Institute of Engineering,
Rua Pedro Nunes-Quinta da Nora, Coimbra, 3030-199, Portugal,
[2] Serviço de Neonatologia, Maternidade Daniel de Matos – ULS Coimbra,
Rua Miguel Torga, 1, 3030-165 Coimbra
[3] RCM2+ Research Centre for Asset Management and Systems Engineering,
Rua Pedro Nunes, Coimbra, 3030-199, Portugal
[*] Corresponding author
a2024161999@isec.pt, mmendes@isec.pt

Abstract

Nutritional adequacy is a key factor in improving the prognosis of very low birth weight infants (VLBW). However, the high workload and scattered clinical data hinder daily systematic analysis. This study investigates the use of locally deployed large language models (LLMs) as assistants for automated data extraction from semi-structured neonatal intensive care unit (NICU) clinical notes. Various text extraction tools were considered (RegEx, ClinicalBERT, MedSpaCy, LLaMA, and Gemma), and an automated analysis pipeline was planned. Then, it is possible that the extracted data can be compared against international neonatal nutritional guidelines (ESPGHAN, SPN) to generate individualized nutritional adjustment proposals. Preliminary results show that LLMs, particularly Gemma 3 12B, can extract clinical data with near-human accuracy and generate context-aware suggestions aligned with clinical practice. This approach offers promising benefits for personalized and efficient nutritional management in VLBW infants, potentially reducing morbidity, mortality, and resource waste.

Keywords: neonatal intensive care; personalized nutrition; LLM; automated data extraction; healthcare artificial intelligence

1 Introduction

Artificial Intelligence has been applied in Medicine with very favorable results, enabling the development and implementation of true precision medicine, with individualized therapeutic and intervention options. In Medicine, despite the large number of models applied in Imaging and in Intensive Care, there is a noticeable gap in the Pediatric and Neonatal areas.

Various Artificial Intelligence models, especially Large Language Models (LLMs), have gained increasing importance as assistants in different fields, as they are capable of quickly absorbing information and answering relatively complex questions in natural language. As an example, we can mention the use of Clinical Key AI©[1], already implemented in several institutions.

In clinical practice, LLMs can offer several advantages, notably assisting professionals in tasks such as identifying similar cases, proposing diagnoses, prognoses, or even therapeutic and nutritional adjustments.

In the present case, the goal is to analyze the feasibility, advantages and limitations of using LLM assistants as support in neonatal intensive care clinical practice, specifically evaluating their capacity to answer

1 https://www.clinicalkey.com/ (last checked on 2025-05-27)

questions in the field of nutrition, assess nutritional practices, and produce adjustment suggestions based on models, consensuses, and guidelines. We aim to improve the prognosis and morbidity/mortality of very low birth weight (VLBW) newborns through better nutritional adaptation to their needs, according to guidelines and consensuses. The model will prepare a nutritional adjustment proposal, based on a comparison between collected parameters and guidelines and consensuses, which the clinician will review and decide whether or not to use.

The population for the study is selected based on Birth Weight (BW) and Gestational Age. Newborns with BW < 1500 g and/or GA ≤ 32 weeks, admitted to the Intensive Care Unit, are selected for the study. The following data will be used: gestational age, day of life, birth weight, current weight, total fluid intake, total caloric intake, total protein intake, total lipid intake, caloric-protein ratio, z-score ratios for weight, length, and head circumference, growth velocity, nutritional supplements used (human milk fortifier, protein supplement), and type of milk used.

If the study results are positive, this model could become a useful tool for daily nutritional adjustment, leading to growth closer to ideal and improving the prognosis and morbidity/mortality of very low birth weight preterm infants. In addition to the health benefits for hospitalized VLBW infants, we also aim to reduce costs by minimizing waste resulting from inadequate and misaligned therapeutic charts.

Ultimately, this model could also be expanded to include other items related to Neonatal Intensive Care.

2 State of the art

Extracting clinical elements from medical texts is a cornerstone of digital health transformation and clinical decision support. Clinical narratives, whether highly structured or entirely free-form, contain the essential details of patient care. In neonatal intensive care units (NICUs), where timely and precise decisions are vital, automating data extraction from clinical notes is not only beneficial but increasingly necessary. This section reviews possible workflows developed to address clinical data extraction, offering a foundation to contextualize the use of large language models in neonatal settings.

2.1 Nature of Clinical Text and Extraction Challenges

Medical texts can be categorized into three levels of organization: structured (*e.g.*, lab results), semi-structured (*e.g.*, clinical progress notes with headings), and unstructured (*e.g.*, free-text consultations or discharge summaries). Structured data can be easily queried, but it often lacks nuance. Unstructured text contains rich clinical context but is difficult to mine systematically.

The extraction of clinical concepts from these texts must account for variability in terminology, abbreviations, negations, temporal references, and context. Medical notes are full of acronyms for which standardization is often lacking, even within a single institution.

2.2 Rule-Based and Pattern-Matching Approaches

Early work relied heavily on rule-based systems, including Regular Expressions (RegEx) and domain-specific scripts. Tools like cTAKES and MetaMap, built on rule-based architectures, map text to standardized medical ontologies like UMLS or SNOMED CT[1], [2]. These systems are designed to identify medical terms in free-text narratives—such as diagnoses, symptoms, or procedures—and align them with structured concepts from ontologies. For example, if a clinical note mentions 'neonatal respiratory distress syndrome', MetaMap can recognize it and link it to its corresponding concept identifier (e.g., C0035222 in UMLS or 69896004 in SNOMED CT). This standardization enables interoperability between systems and supports downstream tasks like analytics, decision support, or population health studies. While powerful in mapping well-known concepts, these tools may struggle

with ambiguous or institution-specific expressions that fall outside predefined vocabularies. These systems offer high precision in tasks such as medication or diagnosis identification but are fragile and require frequent updating to adapt to new text formats or clinical practices. Also, neonatal datasets are scarce.

Pattern-matching remains useful in specific scenarios, particularly when working with templated documentation, such as ICU flowsheets or neonatal daily logs. However, its limitations become apparent when the language used is inconsistent or when semantic understanding is necessary. Context awareness is also an issue.

2.3 Statistical and Machine Learning-Based Systems

The introduction of statistical machine learning improved upon rigid rule-based techniques. Conditional Random Fields (CRFs) and Support Vector Machines (SVMs) became widely used for Named Entity Recognition (NER) in clinical text [3]. These models identify and classify text into predefined clinical categories such as diagnoses, procedures, or medications. For instance, in neonatal care, CRF-based models have been applied to identify 'respiratory support escalation', 'oxygen weaning', or 'necrotizing enterocolitis' within free-text daily progress notes. By learning contextual dependencies, CRFs can disambiguate whether a reference to "ventilation" refers to invasive mechanical ventilation or non-invasive methods like CPAP. Similarly, SVMs trained on annotated corpora can distinguish between therapeutic interventions and diagnostic observations. In neonatal contexts, SVMs have been used experimentally to classify whether phrases such as "NIV escalation" or "oxygen supplementation" represent active interventions versus monitoring status. They have also been tested to differentiate conditions like "intraventricular hemorrhage" as present diagnoses versus historical findings, based on phrase cues. Moreover, these models often struggle to incorporate context beyond sentence level, reducing their utility in extracting temporally and semantically related concepts. These models can be highly effective when trained on domain-specific corpora, but their performance often drops when applied to new datasets, particularly when the clinical language varies significantly across institutions or note types. However, these models often require domain-specific features and substantial amounts of manually labeled data, which are rarely available as neonatal datasets.

2.4 Deep Learning and Contextual Language Models

Deep learning techniques, particularly those based on recurrent neural networks (RNNs) and long short-term memory networks (LSTMs), were a significant step forward. These models offered improved handling of temporal dependencies and non-linear relationships in text. Yet, they were eventually outperformed by transformer-based architectures.

In 2018, Google deployed Bidirectional Encoder Representations from Transformers (BERT). Still today, its clinical adaptations (e.g., ClinicalBERT, BioBERT) continue to offer robust solutions to NER and relation extraction[4], [5]. These models leverage attention mechanisms to consider context from both directions of the input text, providing nuanced understanding that is essential in clinical interpretation. In neonatal domains, these models can extract key information such as type and timing of respiratory support (e.g., 'mechanical ventilation started on D2') or nutritional status (e.g., 'enteral feeds suspended due to abdominal distension'). For example, ClinicalBERT has been applied to identify risk factors in premature infants by capturing subtle cues in documentation that correlate with outcomes like bronchopulmonary dysplasia. The ability to incorporate context allows the model to distinguish between past medical history and current conditions, which is critical in neonatal daily progress notes where temporal awareness is essential. Despite their strengths, such models still require cumbersome fine-tuning to the specific linguistic patterns found in neonatal intensive care notes, as general biomedical corpora may not fully represent neonatal terminology.

For instance, ClinicalBERT has been successfully applied to medication extraction, adverse drug event detection, and de-identification tasks in large datasets like MIMIC-III[6], [7]. MedSpaCy further enhances these capabilities with rule-based extensions, section detection, and negation handling tailored to the clinical domain [8]. It is particularly useful for recognizing the structure and context of clinical documentation. For example, MedSpaCy can identify whether 'apnea episodes noted' appears in the 'Assessment' section versus the 'History' section, changing its interpretation as active versus inactive/past diagnosis. Additionally, its ConText algorithm can determine whether a condition like 'necrotizing enterocolitis' is negated ("no signs of NEC") or hypothetical ("monitor for NEC"). MedSpaCy is especially powerful when used in combination with entity recognizers like spaCy or ClinicalBERT, forming a hybrid rule-based and statistical pipeline that is modular, explainable, and adaptable to local documentation styles.

However, a complete language support is restricted to english (present day), which constitutes a major drawback to effective deployment in this country. The necessary reverse translation is counterintuitive, adding complexity and error-prone.

2.5 Large Language Models in Clinical Practice

The introduction of LLMs like GPT, LLaMA, and PaLM marks a paradigm shift. Unlike earlier models, LLMs can be adapted to diverse tasks through prompting, without needing task-specific training. This flexibility is seen across different families of models: PaLM (Google©) offers multimodal, multitask capabilities integrated into tools like Bard; GPT (OpenAI)© powers applications like ChatGPT and Codex with strong language generation performance; and LLaMA (Meta©) provides efficient, open-source models for local deployment and research. Each has strengths in terms of accessibility, scalability, or domain adaptability, which makes LLMs particularly attractive in low-resource domains such as neonatology, where data are scarce and language diversity requires adaptable solutions [9], [10].

In recent studies, GPT-3 and similar models have shown promising results in summarizing clinical records, generating discharge instructions, and performing zero-shot classification of medical events [10], [11]. Importantly, they can handle temporal and contextual ambiguity more effectively than prior solutions.

However, clinical use of LLMs raises challenges. Hallucination—generation of incorrect but plausible information—is a well-documented issue. Moreover, models may reproduce biases present in training data, and their outputs are not inherently explainable[12]. Clinical deployment thus requires guardrails, such as restricted prompting, human-in-the-loop validation, and domain-specific fine-tuning [13].

2.6 Applications and Case Studies

Several case studies illustrate the successful implementation of clinical NLP tools:

- In oncology, NLP systems have been used to extract staging and treatment plans from pathology reports, enhancing registry completeness and care coordination [14].
- In adult ICU, NLP supports medical decision-making and identifies undiagnosed conditions [15], [16].
- In neonatology, research is more limited. SOFA Score can be predicted from patient notes (also nSOFA) [17]. Preterm birth risk prediction [18] and maternal data extraction are also possible [19].

2.8 Hybrid Pipelines and Integration with EHR Systems

Hybrid pipelines that combine deterministic rules, statistical models, and LLMs are emerging as best practice. For example, a pipeline may use MedSpaCy to detect sections and extract key entities, followed by an LLM to contextualize and summarize the findings. Integration with EHRs could be facilitated by standards such as HL7 FHIR and openEHR, which allow structured outputs to feed clinical dashboards or decision support systems. Presently, in this country, the lack of implementation of such protocols constitutes an irreversible loophole. Also, lack of Portuguese language support is a drawback. Still, theorethically speaking, the former would be the most accurate workflow.

3 Practical Exploration

I) Clinical Data Extraction Models

The most complex and probably least reliable step in the entire process is the extraction of clinical data from text, whether it is highly structured (e.g., medical certificate), semi-structured (e.g., ICU clinical notes), or unstructured/free text (e.g., consultation notes, addenda, etc.).

The available options vary according to the original text, but the analysis is based on the simplest option:

1 Structured Text

a Use of a RegEx (Regular Expression) tool only – as it uses a sequence of characters/search pattern, it can extract specific information. It relies on an exaction pattern. The text would have to be highly structured and rigid. No duplicate data formats are allowed in that text. It would require one RegEx per item intended for extraction.

b Possibly the use of a general LLM could also perform well, but it would require more complex computation.

2 Semi-Structured Text

a Extensive use of RegEx with multiple elements to constrain the extracted data, relying on the most basic structure of the text.

b LLM Models for Element Extraction and Methods for Named Entity Recognition

b.1) BERT / CLINICALBERT – an expanded version of the BERT model more adapted to find medical terms in free text; lighter in computational terms than LLMs; limited to the collection of medical terms; free; not cloud-based; requires the use of MedSpaCy for interaction.

b.2) Lamma x / Gemma x (other) – more generic, more demanding in terms of computing power, less dedicated or specific to medicine; free; not cloud-based; context-sensitive.

c. Custom commercial solutions (such as John Snow Labs©) – paid, dedicated and specific to medicine, customizable on demand (paid), cloud-based

d. Extensive use of RegEx only – it would imply a highly structured text as already stated; one "expression" per item; context-blind.

Table 1 summarizes the advantages and disadvantages of different solutions.

Table 1: Possible pathways – pitfalls and advantages

Solution	Description	Advantages	Pitfalls
RegEx	Uses a search string/pattern Can extract specific information. Implies an exact pattern.	Simple and fast Local; Free	Limited to highly structured texts 2 different data elements must never have similar text Does not contextualize Complex "programming" (?)
BERT/ClinicalBERT	NLP, NER-based Expansion of the BERT model Adapted for medical terms in free/semi-structured text Used for clinical classification, entity recognition extraction, relationship with ontologies (e.g. Snomed) Encoder Tokenization and Vectorization (1) Deterministic	Local; Free Computationally lighter than Lamma / Gemma	Limited to medical term collection only It doesn't work alone: - Complex Programming - Complemented with SpaCy - Complemented by RegEx
Lamma x Gemma x	"Complex", autoregressive NLP "Decoder" Tokenization and Vectorization (2) Probabilistic	Local; Free Can extract Clinical Entities Identifies clinical context and evolution Produces text, tables, etc Adjustment possible with prompt engineering The simplest to integrate into clinical practice (?) No programming required (?) Evolutive (?)	Computationally more demanding Not specific to Medicine Hallucination (can be minimized...) Can be combined with RegEx as well. Prompt Engineering If maximum accuracy is required, precede with ClinicalBert and / or Regex. (3)

Note: (1) fixed vectors per token/phrase for entity extraction (NER), encodes the whole phrase/expression at once. (2) dynamic, contextualized vectors to predict the next token, with inference. (3) extraction is done by ClinicalBERT and the extracted data is processed by Gemma / Lamma, which generates a report, potentially contextualized and interactive.

II) Comparison with Standards

The comparison with standards, aiming at frequent and individualized adjustment of nutritional intake, will be carried out automatically. These standards are available freely and consist, mainly, in the following:

1 *Recommendations for enteral nutrition in the preterm infant* – 2023 update from the Portuguese Neonatology Society:

Part I – Nutritional needs and enteral nutrition during hospitalization
Part II – Specificities of enteral nutrition in special clinical situations and post-discharge feeding

2 *Enteral Nutrition in Preterm Infants: ESPGHAN Position Paper 2022*

These will be summarized and relevant data presented in a table by day of life, gestational age, and birth weight, as well as other relevant somatometric data. A general LLM, possibly the same used previously, will compare the data gathered from the newborn and outline the deviations.

III) Nutritional Intervention Adjustment

The development of an individualized proposal will be carried out by the same LLM model used for data extraction. Eventually, it might be necessary to use another LLM model, more adapted to use the information gathered in steps I and II.

It involves, in a first approach, the preparation of a prompt to systematize the parameters to be compared, the parametrization based on birth weight, gestational age, and day of life, and finally a highly individualized report/proposal. This will be presented to the clinician.

Figure 1 illustrates the workflow proposed.

Workflow

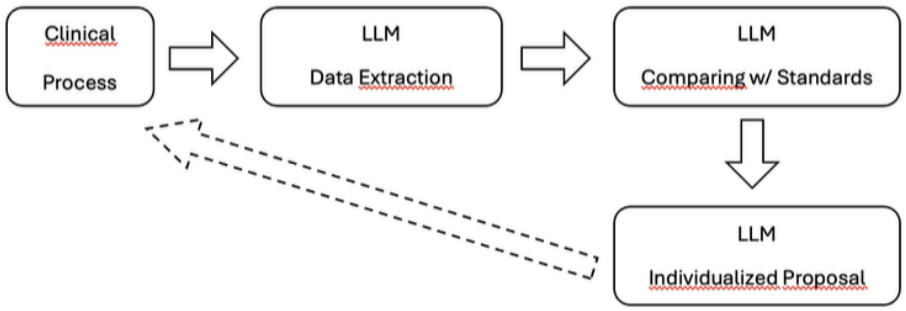


***Figure 1** – Workflow of the method proposed*

Features to Extract and Calculate

Gestational age, day of life, birth weight, current weight, total fluid intake, total caloric intake, total protein intake, total lipid intake, caloric-protein ratio, z-score ratios for weight, length, and head circumference, growth velocity, nutritional supplements in use (human milk fortifier, protein supplement), type of milk used.

4 Methods

A proof-of-concept approach was set up to ascertain the model capabilities to extract relevant data.

The project is based on archetypal clinical records from NICU settings, not real ones, according to patient privacy concerns. Daily progress notes were written in semi-structured format. These notes contain critical information on each infant's nutritional intake, growth parameters, and clinical course.

A pipeline was considered involving several steps:

1 Extract clinical variables using different natural language processing (NLP) strategies - general-purpose LLMs (LLaMA 3.1 4B, Gemma 3 12B) for context-aware data extraction

2 Compare extracted data to the real data in the clinical notes

3 Adjust the NLP parameters to minimize hallucination and grasp the real data

4 Elect the best setup

The models were implemented offline using LM Studio to ensure data security and compliance with hospital data policies. Accuracy and coherence of LLM-generated outputs were assessed against manual expert annotation in test cases. Decoy (non-medical texts) were also used in proofing the LLMs.

5 Results

Among the evaluated methods, Gemma 3 12B showed the best trade-off between performance and system requirements, achieving accurate extraction of key data with minimal hallucination when used with strict prompting. The use of self-consistency techniques (multiple output generation and majority selection) further improved reliability.

The LLM-based system was able to: Automatically identify and structure key clinical features; distinguish between real and decoy clinical notes; map the clinical features according to context; respect real data values.

Some minor improvements are still needed. The collected data need to be validated based on admissible ranges in order to ensure consistency and clinical plausibility (eg the newborn weight in grams must be between 250g and 6000g).

Compared to manual workflows, the system offered significant time savings and reduced variability, especially valuable in busy NICU settings with multiple critical patients.

```
prompt = f"""
Nota para o modelo: AET significa "Aporte Enteral Total". TGV significa "Transfusão de Glóbulos Vermelhos".

Abaixo está um relatório clínico de um recém-nascido.

Por favor, leia cuidadosamente o texto e extraia as seguintes informações clínicas, se estiverem disponíveis:

1. Idade gestacional (em semanas)
2. Peso ao nascimento (em gramas)
3. Dia em que foi iniciada a nutrição enteral
4. Dia em que o recém-nascido atingiu aporte enteral total (150 ml/kg/dia)
5. Tipo de leite utilizado (ex: leite materno, fórmula, leite de banco, fórmula extensamente hidrolisada)
6. Uso de nutrição parenteral (Sim/Não)
7. Se sim, dia de início e término da nutrição parenteral

Por favor, responda em formato estruturado como o exemplo abaixo:

Idade gestacional: ___ semanas
Peso ao nascimento: ___ g
Início da nutrição enteral: Dia ___
Aporte enteral total: Dia ___
Tipo de leite: ___
Nutrição parenteral: Sim/Não
Início da NP: Dia ___
Término da NP: Dia ___

Relatório:
{texto}
"""
```

Figure 1 – Prompt used (in Portuguese) – defining the context and the objective, and the desired structure of the collected data

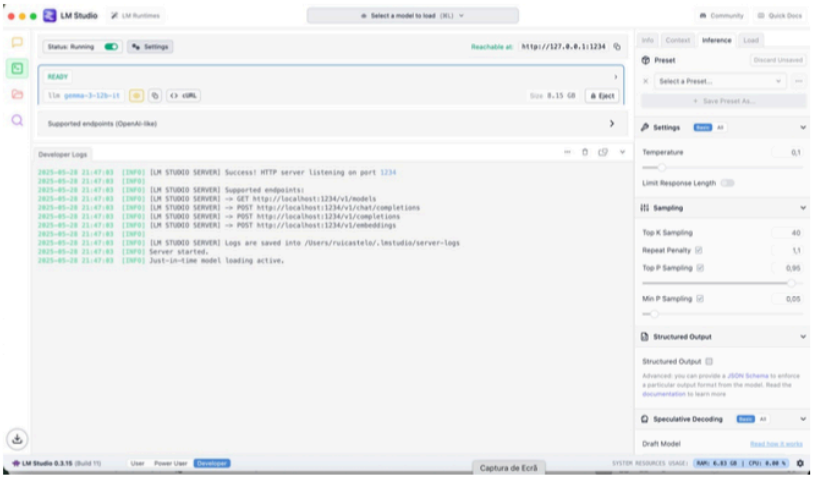


Figure 2 – Using LMStudio to adjust Gemma 3 – topK, topP and temperature

resultados_clinicos_20250527_182302_tabela

Arquivo	Idade gestacional	Peso ao nascimento (g)	Início nutrição enteral (Dia)	Aporte enteral total (Dia)	Tipo de leite	Nutrição parenteral	Início NP (Dia)	Término NP (Dia)
RN1.txt	26	932	7		Leite de banco	Não		
RN3.txt	26	944	9		Fórmula	Não		
RN4.txt	26	1079	8		Fórmula extensamente hidrolisada	Sim		
RN5.txt	26	1001	7		Fórmula extensamente hidrolisada	Sim		
RN6.txt	26	1056	7		Fórmula extensamente hidrolisada	Sim		
RN7.txt	26	1065	8		Fórmula extensamente hidrolisada	Sim		
RN8.txt	26	863	7		Fórmula	Não		
RN9.txt	26	887	8		Não informado	Não		
RN10.txt	26	917	8		Não informado	Não		
RN2.txt	26	1058	7		Fórmula	Não		
RN21.txt		9173469359			Não informado	Sim		
RN22.txt					Não informado	Não		
RN23.txt					Não informado	Não		

Figure 3 – csv table with extracted and structured data (in Portuguese)

6 Discussion

Despite known limitations of LLMs—such as occasional hallucination and non-deterministic behavior—this study shows that strict prompting strategies and offline deployment can mitigate risks and deliver clinically useful outputs. The automated workflow may match or even outperform manual data extraction in terms of speed and consistency, particularly in repetitive tasks. Importantly, the system remains under full clinical control, with LLMs acting as suggestive tools rather than autonomous agents.

This first approach sets the pace for further development of steps II and III

The proposed approach can evolve flexibly: the same pipeline may be adapted to extract additional clinical dimensions (e.g., somatometric data, ventilatory parameters, infection markers) by simply adjusting prompts, without reprogramming or retraining the model. This extensibility makes it suitable for broader integration into clinical decision support systems in neonatal care.

7 Conclusion

Local LLMs show promise as clinical assistants for neonatal nutritional care. Their capacity to extract and process information from semi-structured notes, and to contextualize outputs based on established standards, highlights their potential to support more personalized and efficient care. Although there are some limitations and specific concerns about reliability, this work highlights a non-inferiority level.

Future work will focus on scaling the solution, integrating real-time clinical data, and evaluating long-term outcomes in neonatal growth and morbidity.

References

[1] A. R. Aronson and F. M. Lang, "An overview of MetaMap: Historical perspective and recent advances," *Journal of the American Medical Informatics Association*, vol. 17, no. 3, pp. 229–236, May 2010, doi: 10.1136/jamia.2009.002733.

[2] G. K. Savova *et al.*, "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 507–513, Sep. 2010, doi: 10.1136/jamia.2009.001560.

[3] B. Settles, "Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets." [Online]. Available: http://us.expasy.org/sprot/

[4] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." [Online]. Available: https://github.com/tensorflow/tensor2tensor

[5] E. Alsentzer *et al.*, "Publicly Available Clinical BERT Embeddings," 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/

[6] Y. Wang *et al.*, "Clinical information extraction applications: A literature review," Jan. 01, 2018, *Academic Press Inc.* doi: 10.1016/j.jbi.2017.11.011.

[7] J. Lee *et al.*, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020, doi: 10.1093/bioinformatics/btz682.

[8] H. Eyre *et al.*, "Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python." [Online]. Available: https://spacy.io/

[9] H. Touvron *et al.*, "LLaMA: Open and Efficient Foundation Language Models," Feb. 2023, [Online]. Available: http://arxiv.org/abs/2302.13971

[10] H. Zhou *et al.*, "A Survey of Large Language Models in Medicine: Progress, Application, and Challenge." [Online]. Available: https://github.com/AI-in-Health/MedLLMsPracticalGuide.

[11] Z. Yang *et al.*, "Large Language Model–Based Critical Care Big Data Deployment and Extraction: Descriptive Analysis," *JMIR Med Inform*, vol. 13, 2025, doi: 10.2196/63216.

[12] L. Boonstra *et al.*, *Prompt Engineering*. Google Cloud, 2024.

[13] A. Gatt *et al.*, "From data to text in the neonatal intensive care Unit: Using NLG technology for decision support and information management," *AI Communications*, vol. 22, no. 3, pp. 153–186, 2009, doi: 10.3233/AIC-2009-0453.

[14] S. Abedian *et al.*, "Automated Extraction of Tumor Staging and Diagnosis Information From Surgical Pathology Reports," *JCO Clin Cancer Inform*, vol. 5, pp. 1054–1061, 2021, doi: 10.1200/CCI.21.

[15] E. Urquhart *et al.*, "A pilot feasibility study comparing large language models in extracting key

information from ICU patient text records from an Irish population," *Intensive Care Med Exp*, vol. 12, no. 1, Dec. 2024, doi: 10.1186/s40635-024-00656-1.
[16] W. Zhang *et al.*, "Risk prediction models for intensive care unitacquired weakness in intensive care unit patients: A systematic review," *PLoS One*, vol. 16, no. 9 September, Sep. 2021, doi: 10.1371/journal.pone.0257768.
[17] F. H. Saner, Y. M. Saner, E. Abufarhaneh, D. C. Broering, and D. A. Raptis, "Comparative Analysis of Artificial Intelligence (AI) Languages in Predicting Sequential Organ Failure Assessment (SOFA) Scores," *Cureus*, May 2024, doi: 10.7759/cureus.59662.
[18] L. Sterckx *et al.*, "Clinical information extraction for preterm birth risk prediction," *J Biomed Inform*, vol. 110, Oct. 2020, doi: 10.1016/j.jbi.2020.103544.
[19] S. Abhyankar and D. Demner-Fushman, "A simple method to extract key maternal data from neonatal clinical notes."

Anexo B – Proof of Concept: Abordagem do Problema em UCIN

Com esta proposta pretende-se melhorar a otimização da nutrição de um grupo de risco de prematuros com elevada fragilidade. A automatização deste processo teria amplas vantagens: ultrapassa a quebra e dispersão da atenção numa unidade com vários RN doentes; automatiza vários cálculos; automatiza a extração e comparação da informação; permite elevada velocidade de execução quando comparada com o trabalho manual. Esta perspetiva centra-se na quase ausência de dados disponíveis em formato eletrónico.

A extração de dados através deste modelo é um conceito que vai ser testado. Apesar de poder funcionar em alguns testes e contextos particulares, teria ainda de ser ajustado para reduzir ou eliminar os problemas clássicos dos LLM como a alucinação...

Uma possível linha de raciocínio: apesar dos LLM não serem determinísticos e 100% exatos para extrair dados, a questão é se conseguem ser *pelo menos equivalentes a ou não inferiores* ao desempenho humano. Ou seja, se a extração de dados automática via LLM é equivalente a extração de dados manual, também sujeita a erros, cansaço e distração. Objetivamente, o ganho está na velocidade de desempenho de tarefas e automatização da mesma, mantendo pelo menos uma qualidade equivalente ou não inferior.

Para além do já referido, a expansibilidade é uma vantagem: se para além de dados de aportes nutricionais for julgado necessário recolher dados de somatometria e também compará-los com equivalentes de referência, será que não seria apenas necessário ajustar a(s) prompt(s) ? Este aspecto é importante para a integração na equipa clínica.

Organização geral do projeto

1) Partir do texto – vou usar texto semi-estruturado → diário clínico "sintético" e "realista" de UCIN → ".txt"
2) Usar um LLM para analisar o texto e extrair o que necessito
    Escolher o LLM -> privilegiar soluções locais e não demasiado exigentes
    Enumerar as variáveis que quero extrair
    Criar a prompt para processar os textos tendo em vista a recolha de variáveis
    Exportar em formato universal para utilização múltipla → ".csv"
3) Análise dos dados recolhidos pelo LLM e comparação com os textos originais

Abaixo segue um texto genérico, sintético, realista, pretendendo simular o conteúdo real. Este visa imitar todas as pequenas idiossincrasias, referências e aspetos linguístico, semântico e contextuais.

XXNOMEXX, D30 de vida
IG 26S+2d | IPM 30s+4d
PN 905g | PA 1020g (-15g/24h)

LISTA DE PROBLEMAS:
Prematuridade | EBPN
SDR - NIPPV
A/B
Anemia desde o nascimento (TGV 7/3)
Má progressão ponderal
Hiponatrémia

Aportes: 168mL/Kg/dia (calorias 142kcal/kg/dia, proteinas 4,5g/Kg/dia, lipidos 6.4g/Kg/dia, PER 3.2g/100Kcal, Ferro 5mg/kg) | Na 2.7mEq/Kg/dia
Dispositivos: VNI, SOG

RESPIRATÓRIO: Em NIPPV desde D1, FiO2 21%. Instável com várias A/B, necessidade de O2 e estimulo para recuperar (rastreios septicos neg), mas noção de maior estabilidade após TGV ontem. Sob cafeína oral. GSV em D27 (05/03): pH 7.25, pCO2 51,6 mmHg, HCO3 22,7 mmol/L, EB -45 mmol/L, AG 16,6 mmol/L.

CARDIOVASCULAR: Hemodinamicamente estável. Último lactato D27 (05/03): 5 mmol/L. Ecocardio funcional D29 (7/3): boa função ventriular, FE 45%, Ae/Ao 1,2. FOP. Sem PCA aparente.

RENAL: Diurese regular, não contabilizada. Última função de D27 (05/03): creatinina 0,5 mg/dL e azoto ureico 11mg/dL.

DIGESTIVO: AET desde D8 (14/2), com tolerância. Suplemento proteico desde D20 e FMS desde D21. Dejeções espontâneas e regulares. ABD normal.

METABÓLICO: Normoglicemia desde D20 (26/2). Última GSV em D27: Na 139 mmol/L, K 5 mmol/L, Cl 104 mmol/L, Ca(i) 5.4mg/dL. Sob suplementação com Na+ oral: 2,7meq/Kg/dia. Rastreio DMO-PT em D27 normal.

HEMATOLÓGICO: Anemia desde D0, último hemograma em D27: Hb 8,4g/dL, Htc 24%, Leuc 14100, Plaq 1.019.000/uL, Ret 240.000/uL - decididio em equipa fazer TGV que fez a D29 (7/3), sem intercorrências. Sob ferro oral desde D14 (20/02), atualmente 5mg/Kg/dia.

INFECIOSO: Sem antibioterapia. Último rastreio sético em D27 negativo. Rastreios ESBL seriados negativos, último a 03/03.

NEUROLÓGICO: Tónus e postura adequados à IG. Última EcoTF D21(27/2): sem LMQ. Ligeira hiperecogenecidade do núcleo caudado bilateralmente (menos que o plexo coróide). Doppler da ACA c/ IR de 0.74.

OUTROS: DP nº 1 em D5 (11/02) - normal e nº2 em D14 (20/02). Somatometria D21: peso 875g, comp 34,7cm, PC 22,1cm.

PLANO:
- GSV.
- Somatometria amanhã.

Figura 64 – Exemplo genérico de texto a utilizar (NB – dados fictícios)

Foram criados 10 textos semelhantes a este para serem processados, referenciados como RN1.txt a RN10.txt. Para além destes, foram também criados 3 textos não médicos, um com conteúdo maioritariamente numérico, outro em prosa e um terceiro misto; o objetivo é o de comparar a performance e alucinação ou mismatch com os textos clínicos; estão referenciados como RN21.txt a RN23.txt. Todos estes podem ser encontrados em anexo.

Enumeração de itens a recolher pelo modelo:

Idade gestacional (Semanas + Dias)	Peso ao nascimento (em g)	Início da nutrição enteral (dia de vida)	Aporte enteral total desde (dia de vida)	Tipo de leite LM / LA	Nutrição parenteral (S/N)	Início da NP (dia de vida)	Término da NP (dia de vida)

Após alguma pesquisa, a solução encontrada recorre a algum código python e interfaces gráficos para escolher e interagir com os vários LLMs para poder ajustar parâmetros e ainda um interface gráfico para manipular os textos / .txt
O interface gráfico para ajuste do LLM foi o LMStudio©. O interface gráfico para manipular os ficheiros foi o Gradio©.
Globalmente, o método de extração pode ser sumarizado assim:

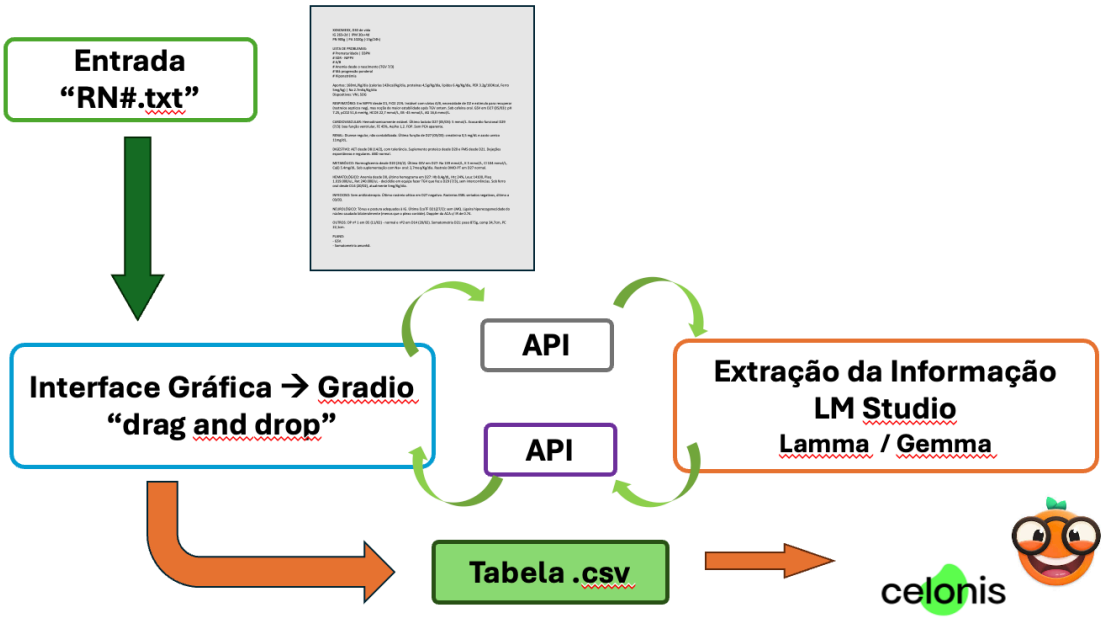


Figura 65 – Fluxo de trabalho esquematizado

```
prompt = f"""
Nota para o modelo: AET significa "Aporte Enteral Total". TGV significa "Transfusão de Glóbulos Vermelhos".

Abaixo está um relatório clínico de um recém-nascido.

Por favor, leia cuidadosamente o texto e extraia as seguintes informações clínicas, se estiverem disponíveis:

1. Idade gestacional (em semanas)
2. Peso ao nascimento (em gramas)
3. Dia em que foi iniciada a nutrição enteral
4. Dia em que o recém-nascido atingiu aporte enteral total (150 ml/kg/dia)
5. Tipo de leite utilizado (ex: leite materno, fórmula, leite de banco, fórmula extensamente hidrolisada)
6. Uso de nutrição parenteral (Sim/Não)
7. Se sim, dia de início e término da nutrição parenteral

Por favor, responda em formato estruturado como o exemplo abaixo:

Idade gestacional: ___ semanas
Peso ao nascimento: ___ g
Início da nutrição enteral: Dia ___
Aporte enteral total: Dia ___
Tipo de leite: ___
Nutrição parenteral: Sim/Não
Início da NP: Dia ___
Término da NP: Dia ___

Relatório:
{texto}
"""
```

Figura 66 – Prompt inserida

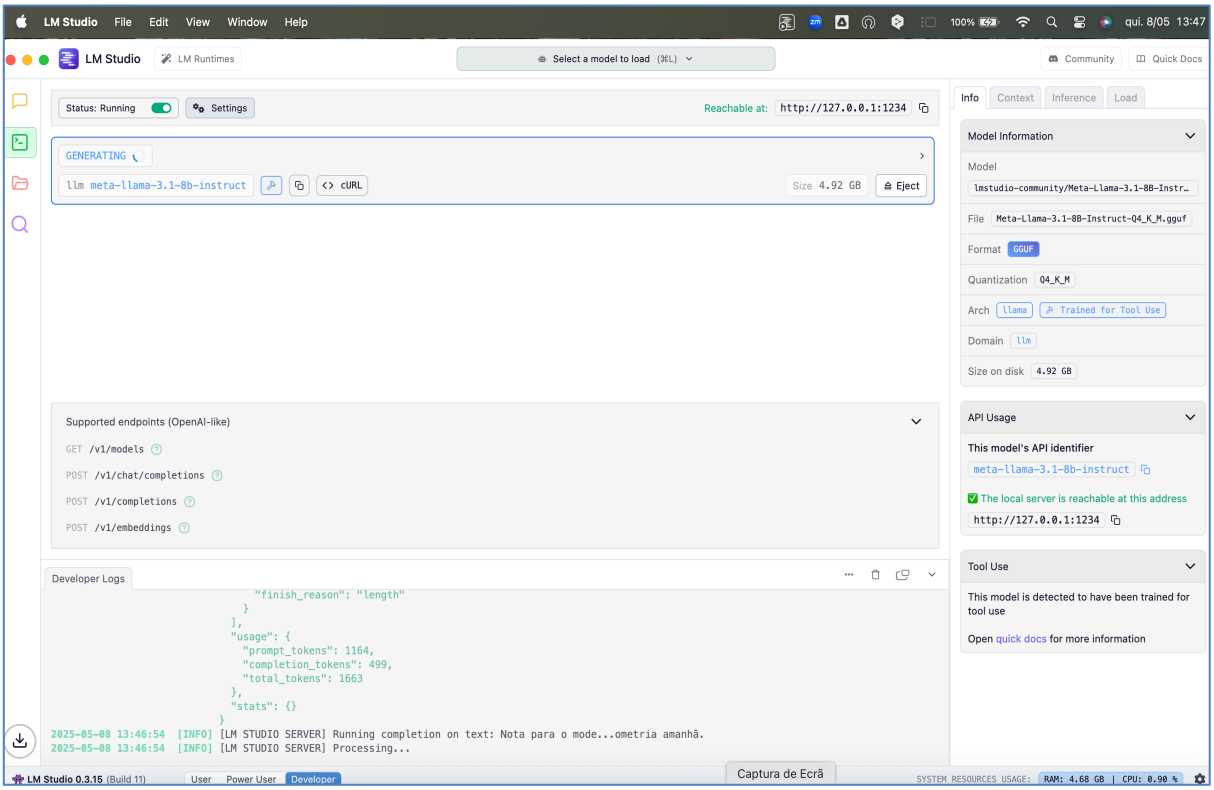


Figura 67 – Interface "gráfico" no LMStudio. Permite ajustar temperatura, top.p e top.k

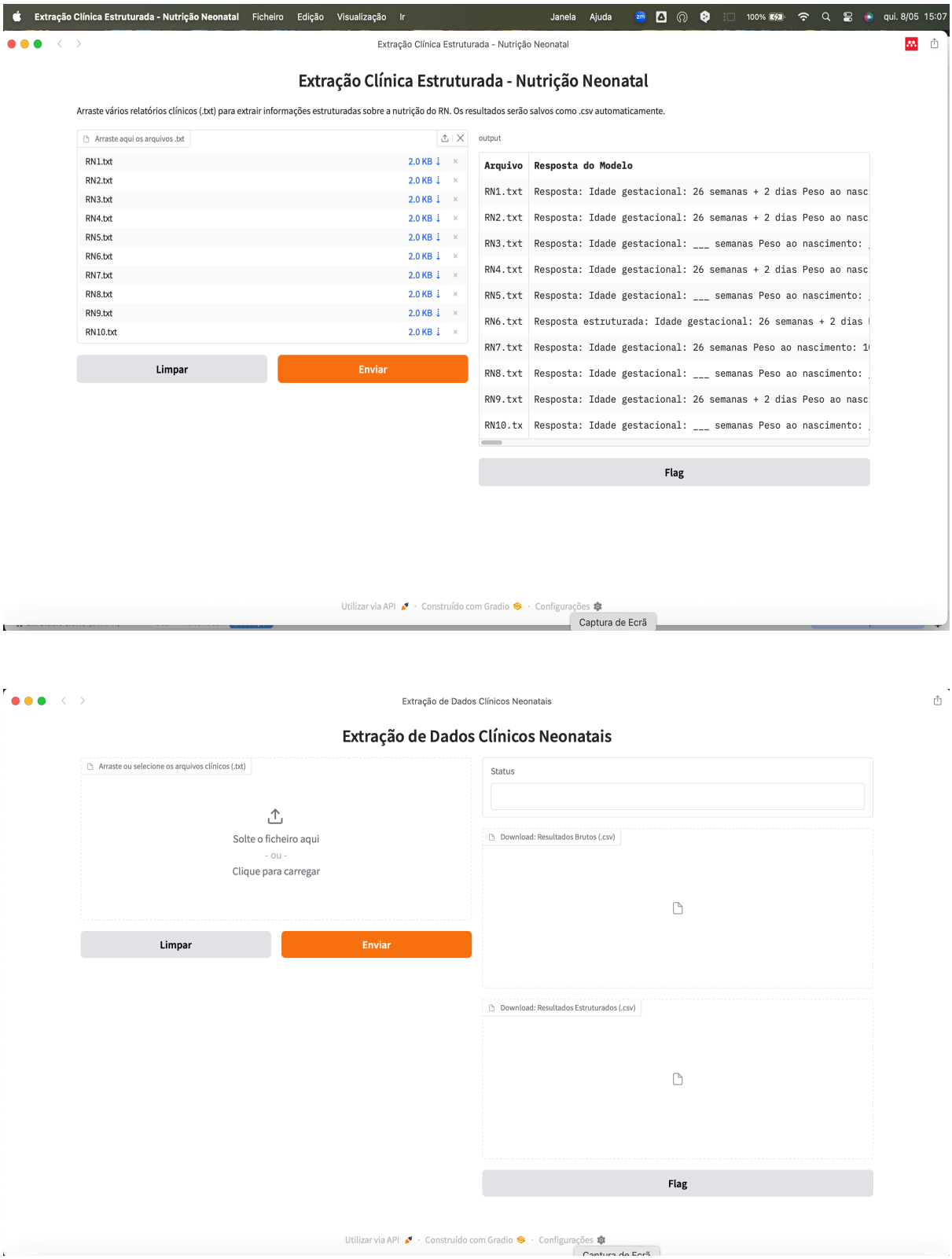
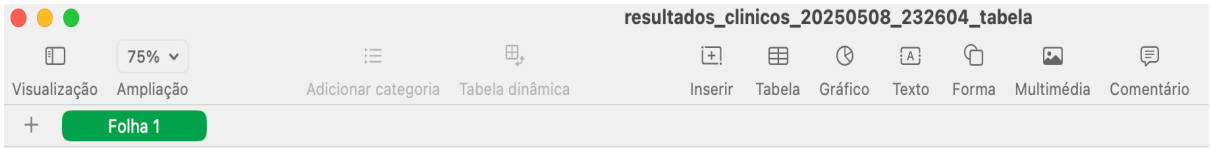


Figura 68 – Interface "gráfico" e Neonatologista-friendly para manipular os textos de diário clínico e, na versão final, permite o download das tabelas .csv

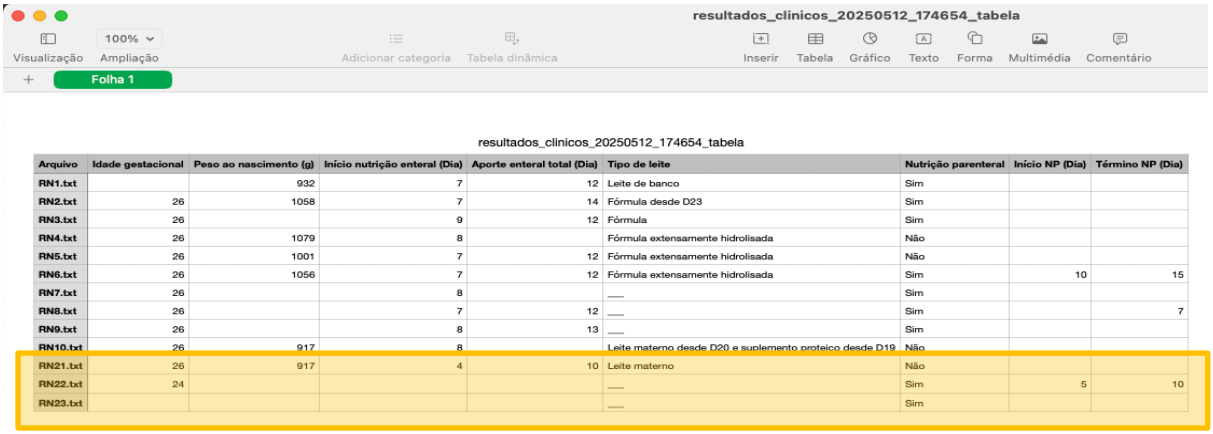


resultados_clinicos_20250508_232604_tabela

Arquivo	Idade gestacional	Peso ao nascimento (g)	Início nutrição enteral (Dia)	Aporte enteral total (Dia)	Tipo de leite	Nutrição parenteral	Início NP (Dia)	Término NP (Dia)
RN1.txt	26	932	7	12	Leite de banco (desde D20) e suplemento proteico (desde D21)	Não		
RN2.txt	26	1058	7	14	Fórmula extensamente hidrolisada	Sim	31	
RN3.txt	26	944	9		Fórmula	Não		
RN4.txt	26	1079	8	32	fórmula extensamente hidrolisada	Sim		
RN5.txt	26		7	28	Fórmula extensamente hidrolisada	Sim	28	
RN6.txt			7	12	fórmula extensamente hidrolisada	Não		
RN7.txt	26	1065	8	29	fórmula extensamente hidrolisada desde D18	Não		
RN8.txt	26	863	7	12	Fórmula	Sim	0	29
RN9.txt	26	887	8	13	Leite materno e suplemento proteico desde D23	Não		
RN10.txt	26	917	8	31	leite materno e suplemento proteico desde D19	Não		

Figura 69 – Tabela csv com os 10 textos sintéticos de diário. Ainda algumas falhas.

Depois tentou-se a integração de 3 textos "*decoy*" – codificados com RN2x.txt.
Estes continham um trajecto gerado por IA com múltiplos números e texto; uma receita culinária com texto e números em gramas e um resumo do Memorial do Convento criada por IA.



resultados_clinicos_20250512_174654_tabela

Arquivo	Idade gestacional	Peso ao nascimento (g)	Início nutrição enteral (Dia)	Aporte enteral total (Dia)	Tipo de leite	Nutrição parenteral	Início NP (Dia)	Término NP (Dia)
RN1.txt		932	7	12	Leite de banco	Sim		
RN2.txt	26	1058	7	14	Fórmula desde D23	Sim		
RN3.txt	26		9	12	Fórmula	Sim		
RN4.txt	26	1079	8		Fórmula extensamente hidrolisada	Não		
RN5.txt	26	1001	7	12	Fórmula extensamente hidrolisada	Não		
RN6.txt	26	1056	7	12	Fórmula extensamente hidrolisada	Sim	10	15
RN7.txt	26		8		___	Sim		
RN8.txt	26		7	12	___	Sim		7
RN9.txt	26		8	13	___	Sim		
RN10.txt	26	917	8		Leite materno desde D20 e suplemento proteico desde D19	Não		
RN21.txt	26	917	4	10	Leite materno	Não		
RN22.txt	24				___	Sim	5	10
RN23.txt					___	Sim		

Figura 70 – Utilização de textos não médicos (*decoy)*

A versão inicial não teve um bom desempenho...

A utilização dos "*decoy*" gerou alguns dados inexistentes... Assim, foram necessários alguns ajustes ao longo das várias versões:

→ Ajustar alguns parâmetros no Lamma: Temperatura 0,8 → 0,7 → 0,6 (reduzir a criatividade)

→ Experimentei o Lamma 3,1 - 4B (RAM 5,6/16Gb) ✅ → Gemma 3 - 12B (RAM 8,3/16Gb) ✅

→ Tentei também o Granite 8B (❌) mas revelou-se menos preciso

→ Austes a Prompt – menos instruções, indicação para ser mais estrita...

→ Tentei incluir Self-Consistency – **V4** gera 3 respostas e depois elege a mais frequente

→ O Gemma 3 funciona por defeito com temperatura 0,1

A versão final condensa todas as alterações e parece a mais funcional.

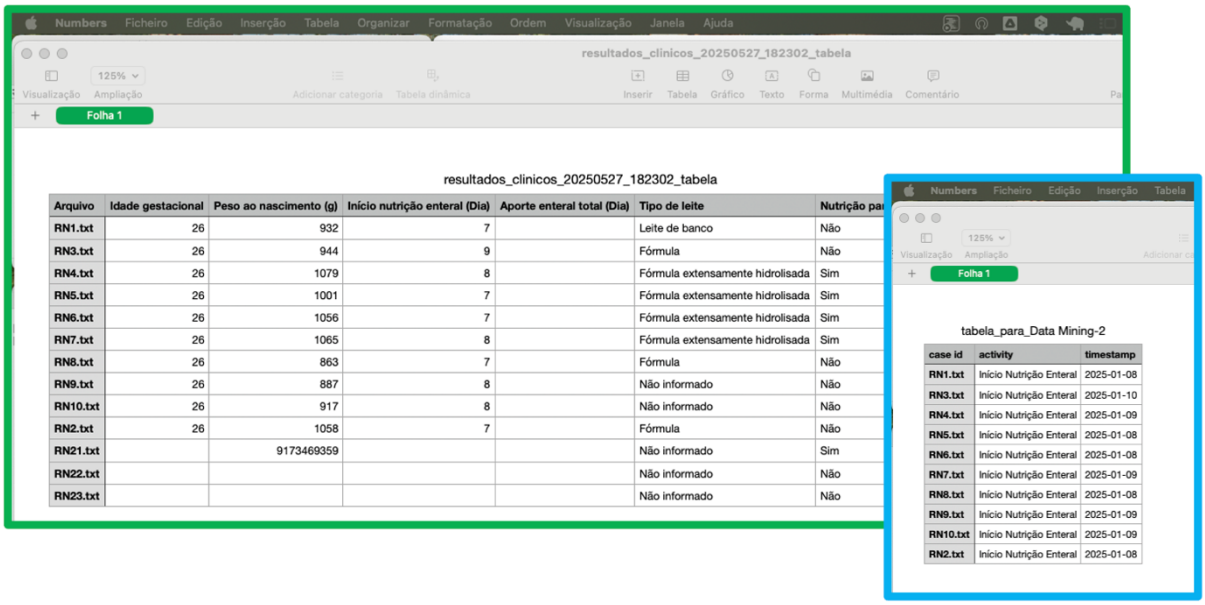


Figura 71 – Tabela csv e tabela com time-stamp geradas pela versão final

A versão final, apesar de não ser perfeita, parece servir para testar o proof-of-concept, aparentemente com resultados favoráveis.

Foram testados 3 modelos de LLM. A escolha destes foi influenciada por 2 aspetos fundamentais:

1) O Projeto teria de funcionar totalmente offline, por uma questão de segurança de dados clínicos (também fundamental no Projeto da Tese);
2) O envio de dados para processamento na cloud, após anonimização dos dados já foi utilizado noutros trabalhos (REF), mas iria tornar o modelo final do Projeto de Tese mais complexo e pesado, e não isento de risco, apesar de tudo;
3) A limitação de hardware – utilizei um MacBook Air M3 com 16 Gb de RAM.
4) O objetivo final é desenvolver um projeto que possa funcionar num vulgar terminal existente na maioria das Instituições de Saúde.

Em face destas premissas, foram testados apenas três modelos:

a) Lamma 3,1 - 4B (RAM 5,6/16Gb)
b) Granite 8B (RAM 6,5/16 Gb)
c) Gemma 3 - 12B (RAM 8,3/16Gb)

Todos eles funcionaram adequadamente via LMStudio©, permitindo os ajustes já referidos, e processaram todos os textos. Globalmente, o Gemma 3 com 12 B foi o melhor a recolher os dados; o Granite foi que apresentou mais erros.

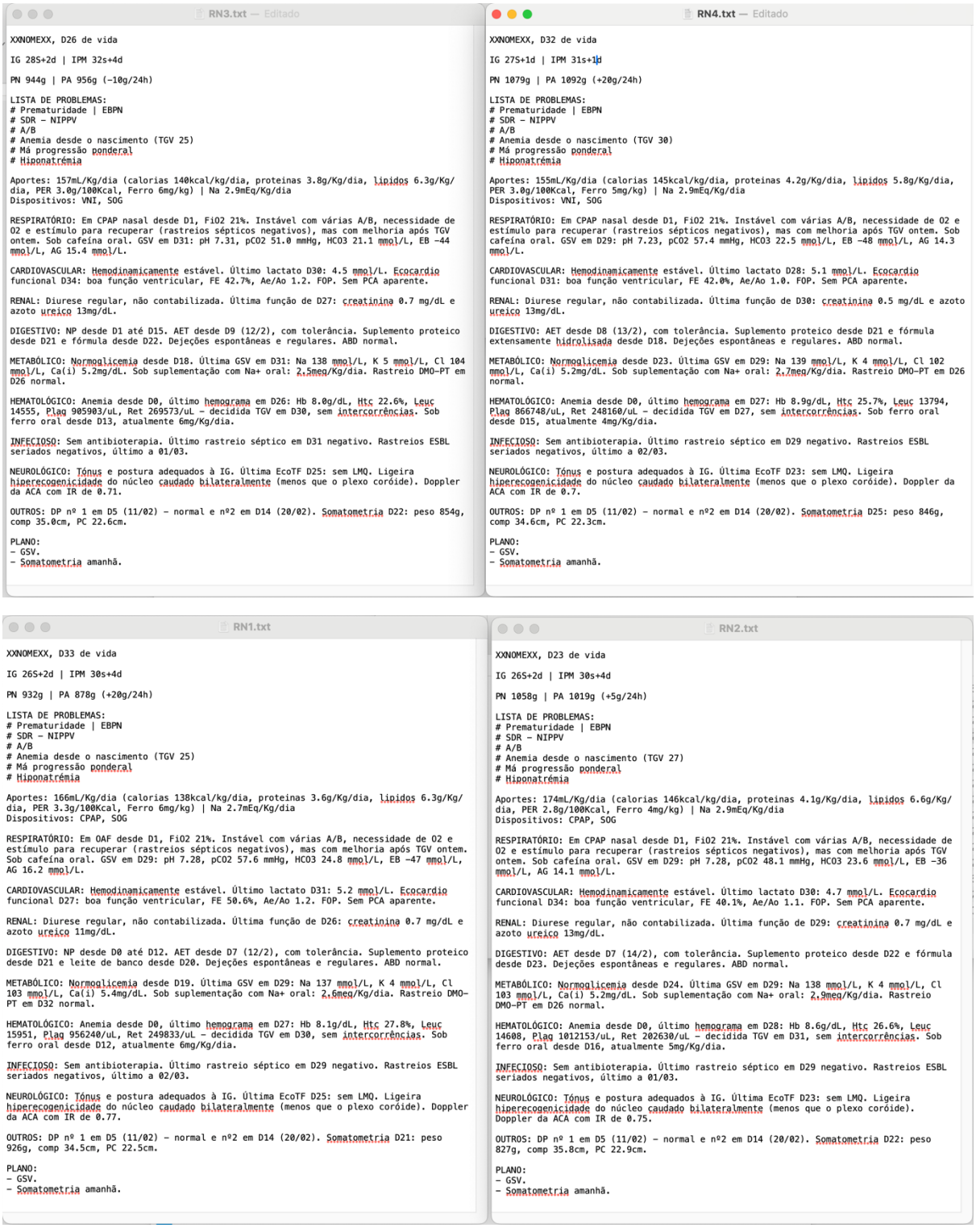
Discussão e Comentários:

Este *Proof of Concept* permitiu evidenciar a capacidade de um LLM extrair dados de textos clínicos com exatidão, apesar de ser em modo "*one shot". Não foram avaliadas medidas como accuracy, recall, precision, F1-score; Krippendorff's alfa e métricas de erro.*

Visto todo o projeto ter funcionado offline, a segurança dos dados clínicos e a não exposição de todo o sistema ao exterior estão assegurados. Também o fato de todo o sistema ter funcionado com um modelo *"opensource"* e final, sem necessitar de aprendizagem, ajuda à simplicidade que se pretende.

Por último, temos a possibilidade de ajustar os dados a extrair fazendo algumas alterações na *prompt*, uma mais-valia para o Projeto de Tese.

Anexo C – Exemplos dos Textos Clínicos do Dataset Sintético



RN3.txt — Editado

```
XXNOMEXX, D26 de vida

IG 28S+2d | IPM 32s+4d

PN 944g | PA 956g (-10g/24h)

LISTA DE PROBLEMAS:
# Prematuridade | EBPN
# SDR - NIPPV
# A/B
# Anemia desde o nascimento (TGV 25)
# Má progressão ponderal
# Hiponatrémia

Aportes: 157mL/Kg/dia (calorias 140kcal/kg/dia, proteinas 3.8g/Kg/dia, lipidos 6.3g/Kg/
dia, PER 3.0g/100Kcal, Ferro 6mg/kg) | Na 2.9mEq/Kg/dia
Dispositivos: VNI, SOG

RESPIRATÓRIO: Em CPAP nasal desde D1, FiO2 21%. Instável com várias A/B, necessidade de
O2 e estímulo para recuperar (rastreios sépticos negativos), mas com melhoria após TGV
ontem. Sob cafeína oral. GSV em D31: pH 7.31, pCO2 51.0 mmHg, HCO3 21.1 mmol/L, EB -44
mmol/L, AG 15.4 mmol/L.

CARDIOVASCULAR: Hemodinamicamente estável. Último lactato D30: 4.5 mmol/L. Ecocardio
funcional D34: boa função ventricular, FE 42.7%, Ae/Ao 1.2. FOP. Sem PCA aparente.

RENAL: Diurese regular, não contabilizada. Última função de D27: creatinina 0.7 mg/dL e
azoto ureico 13mg/dL.

DIGESTIVO: NP desde D1 até D15. AET desde D9 (12/2), com tolerância. Suplemento proteico
desde D21 e fórmula desde D22. Dejeções espontâneas e regulares. ABD normal.

METABÓLICO: Normoglicemia desde D18. Última GSV em D31: Na 138 mmol/L, K 5 mmol/L, Cl 104
mmol/L, Ca(i) 5.2mg/dL. Sob suplementação com Na+ oral: 2.5meq/Kg/dia. Rastreio DMO-PT em
D26 normal.

HEMATOLÓGICO: Anemia desde D0, último hemograma em D26: Hb 8.0g/dL, Htc 22.6%, Leuc
14555, Plaq 905903/uL, Ret 269573/uL - decidida TGV em D30, sem intercorrências. Sob
ferro oral desde D13, atualmente 6mg/Kg/dia.

INFECIOSO: Sem antibioterapia. Último rastreio séptico em D31 negativo. Rastreios ESBL
seriados negativos, último a 01/03.

NEUROLÓGICO: Tónus e postura adequados à IG. Última EcoTF D25: sem LMQ. Ligeira
hiperecogenicidade do núcleo caudado bilateralmente (menos que o plexo coróide). Doppler
da ACA com IR de 0.71.

OUTROS: DP nº 1 em D5 (11/02) - normal e nº2 em D14 (20/02). Somatometria D22: peso 854g,
comp 35.0cm, PC 22.6cm.

PLANO:
- GSV.
- Somatometria amanhã.
```

RN4.txt — Editado

```
XXNOMEXX, D32 de vida

IG 27S+1d | IPM 31s+1d

PN 1079g | PA 1092g (+20g/24h)

LISTA DE PROBLEMAS:
# Prematuridade | EBPN
# SDR - NIPPV
# A/B
# Anemia desde o nascimento (TGV 30)
# Má progressão ponderal
# Hiponatrémia

Aportes: 155mL/Kg/dia (calorias 145kcal/kg/dia, proteinas 4.2g/Kg/dia, lipidos 5.8g/Kg/dia,
PER 3.0g/100Kcal, Ferro 5mg/kg) | Na 2.9mEq/Kg/dia
Dispositivos: VNI, SOG

RESPIRATÓRIO: Em CPAP nasal desde D1, FiO2 21%. Instável com várias A/B, necessidade de O2 e
estímulo para recuperar (rastreios sépticos negativos), mas com melhoria após TGV ontem. Sob
cafeína oral. GSV em D29: pH 7.23, pCO2 57.4 mmHg, HCO3 22.5 mmol/L, EB -48 mmol/L, AG 14.3
mmol/L.

CARDIOVASCULAR: Hemodinamicamente estável. Último lactato D28: 5.1 mmol/L. Ecocardio
funcional D31: boa função ventricular, FE 42.0%, Ae/Ao 1.0. FOP. Sem PCA aparente.

RENAL: Diurese regular, não contabilizada. Última função de D30: creatinina 0.5 mg/dL e azoto
ureico 13mg/dL.

DIGESTIVO: AET desde D8 (13/2), com tolerância. Suplemento proteico desde D21 e fórmula
extensamente hidrolisada desde D18. Dejeções espontâneas e regulares. ABD normal.

METABÓLICO: Normoglicemia desde D23. Última GSV em D29: Na 139 mmol/L, K 4 mmol/L, Cl 102
mmol/L, Ca(i) 5.2mg/dL. Sob suplementação com Na+ oral: 2.7meq/Kg/dia. Rastreio DMO-PT em D26
normal.

HEMATOLÓGICO: Anemia desde D0, último hemograma em D27: Hb 8.9g/dL, Htc 25.7%, Leuc 13794,
Plaq 866748/uL, Ret 248160/uL - decidida TGV em D27, sem intercorrências. Sob ferro oral
desde D15, atualmente 4mg/Kg/dia.

INFECIOSO: Sem antibioterapia. Último rastreio séptico em D29 negativo. Rastreios ESBL
seriados negativos, último a 02/03.

NEUROLÓGICO: Tónus e postura adequados à IG. Última EcoTF D23: sem LMQ. Ligeira
hiperecogenicidade do núcleo caudado bilateralmente (menos que o plexo coróide). Doppler da
ACA com IR de 0.7.

OUTROS: DP nº 1 em D5 (11/02) - normal e nº2 em D14 (20/02). Somatometria D25: peso 846g,
comp 34.6cm, PC 22.3cm.

PLANO:
- GSV.
- Somatometria amanhã.
```

RN1.txt

```
XXNOMEXX, D33 de vida

IG 26S+2d | IPM 30s+4d

PN 932g | PA 878g (+20g/24h)

LISTA DE PROBLEMAS:
# Prematuridade | EBPN
# SDR - NIPPV
# A/B
# Anemia desde o nascimento (TGV 25)
# Má progressão ponderal
# Hiponatrémia

Aportes: 166mL/Kg/dia (calorias 138kcal/kg/dia, proteinas 3.6g/Kg/dia, lipidos 6.3g/Kg/
dia, PER 3.3g/100Kcal, Ferro 6mg/kg) | Na 2.7mEq/Kg/dia
Dispositivos: CPAP, SOG

RESPIRATÓRIO: Em OAF desde D1, FiO2 21%. Instável com várias A/B, necessidade de O2 e
estímulo para recuperar (rastreios sépticos negativos), mas com melhoria após TGV ontem.
Sob cafeína oral. GSV em D29: pH 7.28, pCO2 57.6 mmHg, HCO3 24.8 mmol/L, EB -47 mmol/L,
AG 16.2 mmol/L.

CARDIOVASCULAR: Hemodinamicamente estável. Último lactato D31: 5.2 mmol/L. Ecocardio
funcional D27: boa função ventricular, FE 50.6%, Ae/Ao 1.2. FOP. Sem PCA aparente.

RENAL: Diurese regular, não contabilizada. Última função de D26: creatinina 0.7 mg/dL e
azoto ureico 11mg/dL.

DIGESTIVO: NP desde D0 até D12. AET desde D7 (12/2), com tolerância. Suplemento proteico
desde D21 e leite de banco desde D20. Dejeções espontâneas e regulares. ABD normal.

METABÓLICO: Normoglicemia desde D19. Última GSV em D29: Na 137 mmol/L, K 4 mmol/L, Cl
103 mmol/L, Ca(i) 5.4mg/dL. Sob suplementação com Na+ oral: 2.6meq/Kg/dia. Rastreio DMO-
PT em D32 normal.

HEMATOLÓGICO: Anemia desde D0, último hemograma em D27: Hb 8.1g/dL, Htc 27.8%, Leuc
15951, Plaq 956240/uL, Ret 249833/uL - decidida TGV em D30, sem intercorrências. Sob
ferro oral desde D12, atualmente 6mg/Kg/dia.

INFECIOSO: Sem antibioterapia. Último rastreio séptico em D29 negativo. Rastreios ESBL
seriados negativos, último a 02/03.

NEUROLÓGICO: Tónus e postura adequados à IG. Última EcoTF D25: sem LMQ. Ligeira
hiperecogenicidade do núcleo caudado bilateralmente (menos que o plexo coróide). Doppler
da ACA com IR de 0.77.

OUTROS: DP nº 1 em D5 (11/02) - normal e nº2 em D14 (20/02). Somatometria D21: peso
926g, comp 34.5cm, PC 22.5cm.

PLANO:
- GSV.
- Somatometria amanhã.
```

RN2.txt

```
XXNOMEXX, D23 de vida

IG 26S+2d | IPM 30s+4d

PN 1058g | PA 1019g (+5g/24h)

LISTA DE PROBLEMAS:
# Prematuridade | EBPN
# SDR - NIPPV
# A/B
# Anemia desde o nascimento (TGV 27)
# Má progressão ponderal
# Hiponatrémia

Aportes: 174mL/Kg/dia (calorias 146kcal/kg/dia, proteinas 4.1g/Kg/dia, lipidos 6.6g/Kg/
dia, PER 2.8g/100Kcal, Ferro 4mg/kg) | Na 2.9mEq/Kg/dia
Dispositivos: CPAP, SOG

RESPIRATÓRIO: Em CPAP nasal desde D1, FiO2 21%. Instável com várias A/B, necessidade de
O2 e estímulo para recuperar (rastreios sépticos negativos), mas com melhoria após TGV
ontem. Sob cafeína oral. GSV em D29: pH 7.28, pCO2 48.1 mmHg, HCO3 23.6 mmol/L, EB -36
mmol/L, AG 14.1 mmol/L.

CARDIOVASCULAR: Hemodinamicamente estável. Último lactato D30: 4.7 mmol/L. Ecocardio
funcional D34: boa função ventricular, FE 40.1%, Ae/Ao 1.1. FOP. Sem PCA aparente.

RENAL: Diurese regular, não contabilizada. Última função de D29: creatinina 0.7 mg/dL e
azoto ureico 13mg/dL.

DIGESTIVO: AET desde D7 (14/2), com tolerância. Suplemento proteico desde D22 e fórmula
desde D23. Dejeções espontâneas e regulares. ABD normal.

METABÓLICO: Normoglicemia desde D24. Última GSV em D29: Na 138 mmol/L, K 4 mmol/L, Cl
103 mmol/L, Ca(i) 5.2mg/dL. Sob suplementação com Na+ oral: 2.9meq/Kg/dia. Rastreio
DMO-PT em D26 normal.

HEMATOLÓGICO: Anemia desde D0, último hemograma em D28: Hb 8.6g/dL, Htc 26.6%, Leuc
14608, Plaq 1012153/uL, Ret 202630/uL - decidida TGV em D31, sem intercorrências. Sob
ferro oral desde D16, atualmente 5mg/Kg/dia.

INFECIOSO: Sem antibioterapia. Último rastreio séptico em D29 negativo. Rastreios ESBL
seriados negativos, último a 01/03.

NEUROLÓGICO: Tónus e postura adequados à IG. Última EcoTF D23: sem LMQ. Ligeira
hiperecogenicidade do núcleo caudado bilateralmente (menos que o plexo coróide).
Doppler da ACA com IR de 0.75.

OUTROS: DP nº 1 em D5 (11/02) - normal e nº2 em D14 (20/02). Somatometria D22: peso
827g, comp 35.8cm, PC 22.9cm.

PLANO:
- GSV.
- Somatometria amanhã.
```

Anexo D – Referenciais de Z-Scores e Aportes Nutricionais

As tabelas de Z-Scores são baseadas nos standards Intergrowth-21st

Lancet 2016: doi.org/10.1016/S0140--6736(16)00384--6.

The International Very Preterm Size at Birth Reference Charts

UNIVERSITY OF OXFORD

Weight (kg) Girls

Gestational age (weeks+days)	Centiles						
	3rd	5th	10th	50th	90th	95th	97th
24+0	0.42	0.44	0.47	0.60	0.77	0.83	0.87
24+1	0.43	0.45	0.48	0.61	0.79	0.84	0.88
24+2	0.44	0.46	0.49	0.63	0.80	0.86	0.90
24+3	0.44	0.46	0.50	0.64	0.82	0.88	0.92
24+4	0.45	0.47	0.51	0.65	0.83	0.89	0.94
24+5	0.46	0.48	0.52	0.66	0.85	0.91	0.95
24+6	0.47	0.49	0.53	0.68	0.87	0.93	0.97
25+0	0.48	0.50	0.54	0.69	0.88	0.95	0.99
25+1	0.49	0.51	0.55	0.70	0.90	0.96	1.01
25+2	0.50	0.52	0.56	0.71	0.92	0.98	1.03
25+3	0.51	0.53	0.57	0.73	0.93	1.00	1.05
25+4	0.52	0.54	0.58	0.74	0.95	1.02	1.07
25+5	0.53	0.55	0.59	0.76	0.97	1.04	1.09
25+6	0.54						
26+0	0.55						
26+1	0.56						
26+2	0.57						

sexo	medida	idade_pma_semanas	fonte	z	valor	unidade
Feminino	comprimento	24.0	muito_pré-termo_nascimento	-3	24.0	cm
Feminino	comprimento	24.0	muito_pré-termo_nascimento	-2	26.6	cm
Feminino	comprimento	24.0	muito_pré-termo_nascimento	-1	29.2	cm
Feminino	comprimento	24.0	muito_pré-termo_nascimento	0	31.8	cm
Feminino	comprimento	24.0	muito_pré-termo_nascimento	1	34.4	cm
Feminino	comprimento	24.0	muito_pré-termo_nascimento	2	37.0	cm
Feminino	comprimento	24.0	muito_pré-termo_nascimento	3	39.6	cm
Feminino	comprimento	24.1429	muito_pré-termo_nascimento	-3	24.2	cm
Feminino	comprimento	24.1429	muito_pré-termo_nascimento	-2	26.8	cm
Feminino	comprimento	24.1429	muito_pré-termo_nascimento	-1	29.4	cm
Feminino	comprimento	24.1429	muito_pré-termo_nascimento	0	32.0	cm
Feminino	comprimento	24.1429	muito_pré-termo_nascimento	1	34.6	cm
Feminino	comprimento	24.1429	muito_pré-termo_nascimento	2	37.2	cm
Feminino	comprimento	24.1429	muito_pré-termo_nascimento	3	39.8	cm
Feminino	comprimento	24.2857	muito_pré-termo_nascimento	-3	24.4	cm
Feminino	comprimento	24.2857	muito_pré-termo_nascimento	-2	27.0	cm
Feminino	comprimento	24.2857	muito_pré-termo_nascimento	-1	29.6	cm
Feminino	comprimento	24.2857	muito_pré-termo_nascimento	0	32.2	cm
Feminino	comprimento	24.2857	muito_pré-termo_nascimento	1	34.8	cm
Feminino	comprimento	24.2857	muito_pré-termo_nascimento	2	37.4	cm
Feminino	comprimento	24.2857	muito_pré-termo_nascimento	3	39.9	cm

As tabelas de Aportes foram baseadas nos Consensos de Nutrição da Soc. Port. de Neonatologia.

Pereira-Da-silva, L., Pissarra, S., Gomes, A., Barroso, R., Fernandes, C., Virella, D., Macedo, I., Santos, E., Teles, A., & Cardoso, M. (2023). Guidelines for enteral nutrition in infants born preterm: 2023 update by the Portuguese Neonatal Society. Part I. Nutrient requirements and enteral feeding approach during the hospital stay. *Portuguese Journal of Pediatrics*, *54*(4), 253–263. https://doi.org/10.24875/PJP.23000004

Pereira-Da-silva, L., Pissarra, S., Gomes, A., Barroso, R., Fernandes, C., Virella, D., Macedo, I., Santos, E., Teles, A., & Cardoso, M. (2023). Guidelines for enteral nutrition in infants born preterm: 2023 update by the Portuguese Neonatal Society. Part II. Enteral feeding in specific clinical conditions and feeding after discharge. *Portuguese Journal of Pediatrics*, *54*(4), 264–270. https://doi.org/10.24875/PJP.23000005

Aportes

Variavel	Unidade	Regra	Limite_Inferior	Limite_Superior	Quando_Aplicar_BW	Quando_Aplicar_Cond
Aporte hídrico total	mL/kg/d	Intervalo	150.0	180.0	BW 0–999 g	Estável
Aporte hídrico total	mL/kg/d	Intervalo	135.0	150.0	BW 0–999 g	DBP/PCA
Aporte hídrico total	mL/kg/d	Max		200.0	BW 0–999 g	LM não fortificado
Aporte hídrico total	mL/kg/d	Intervalo	150.0	180.0	BW 1000–1499 g	Estável
Aporte hídrico total	mL/kg/d	Intervalo	135.0	150.0	BW 1000–1499 g	DBP/PCA
Aporte hídrico total	mL/kg/d	Max		200.0	BW 1000–1499 g	LM não fortificado
Aporte hídrico total	mL/kg/d	Intervalo	150.0	180.0	BW 1500–2499 g	Estável
Aporte hídrico total	mL/kg/d	Intervalo	135.0	150.0	BW 1500–2499 g	DBP/PCA
Aporte hídrico total	mL/kg/d	Max		200.0	BW 1500–2499 g	LM não fortificado
Energia total	kcal/kg/d	Intervalo	115.0	140.0	BW 0–999 g	Estável
Energia total	kcal/kg/d	Intervalo	120.0	150.0	BW 0–999 g	DBP
Energia total	kcal/kg/d	Intervalo	140.0	160.0	BW 0–999 g	Crescimento subótimo
Energia total	kcal/kg/d	Intervalo	115.0	140.0	BW 1000–1499 g	Estável
Energia total	kcal/kg/d	Intervalo	120.0	150.0	BW 1000–1499 g	DBP
Energia total	kcal/kg/d	Intervalo	140.0	160.0	BW 1000–1499 g	Crescimento subótimo
Energia total	kcal/kg/d	Intervalo	115.0	140.0	BW 1500–2499 g	Estável
Energia total	kcal/kg/d	Intervalo	120.0	150.0	BW 1500–2499 g	DBP
Energia total	kcal/kg/d	Intervalo	140.0	160.0	BW 1500–2499 g	Crescimento subótimo
Proteína	g/kg/d	Intervalo	3.5	4.0	BW 0–999 g	Estável
Proteína	g/kg/d	Max		4.5	BW 0–999 g	Crescimento subótimo (se ureia <34 mg/dL)
Proteína	g/kg/d	Intervalo	3.5	4.0	BW 1000–1499 g	Estável
Proteína	g/kg/d	Max		4.5	BW 1000–1499 g	Crescimento subótimo (se ureia <34 mg/dL)
Proteína	g/kg/d	Intervalo	3.5	4.0	BW 1500–2499 g	Estável
Proteína	g/kg/d	Max		4.5	BW 1500–2499 g	Crescimento subótimo (se ureia <34 mg/dL)

Instituto Superior
de Engenharia
Politécnico de Coimbra