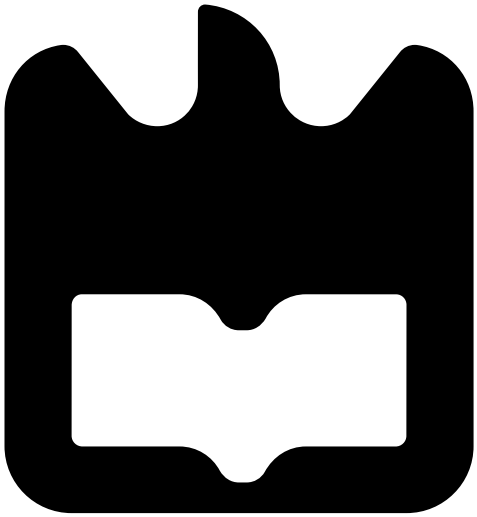

# Relatório do Trabalho Laboratorial nº 3

Informação e Codificação (2025/26)

**Pedro Miguel Miranda de Melo** (114208)
**Rúben Cardeal Costa** (114190)
**Hugo Marques Dias** (114142)
*Departamento de Eletrónica, Telecomunicações e Informática (DETI)*
*Universidade de Aveiro*

Novembro de 2025

# Conteúdo

# 1 Introdução

O presente relatório técnico descreve o trabalho realizado no âmbito do projeto de compressão de um Grande Modelo de Linguagem (LLM). O objetivo central passa por desenvolver uma estratégia de compressão otimizada e eficiente para o ficheiro `model.safetensors` ($\sim 1$ GB) que contém os parâmetros de um LLM. Continuar...

# 2 Análise e Caracterização da Fonte

Para desenhar um codec eficiente, é imperativo compreender a natureza estatística da fonte de informação. Esta secção detalha a análise teórica e experimental realizada sobre o ficheiro `model.safetensors`.

## 2.1 Análise Estrutural (Formato de Dados)

A inspeção do cabeçalho do ficheiro revelou que os dados estão armazenados no formato **BF16** (*Brain Floating Point 16*). Ao contrário de inteiros de 16 bits, onde a distribuição de bits tende a ser uniforme em dados aleatórios, o BF16 possui uma semântica específica composta por:

- **1 bit de Sinal ($S$)** e **8 bits de Expoente ($E$)**: Ocupam maioritariamente o byte mais significativo (MSB).
- **7 bits de Mantissa ($M$)**: Ocupam o byte menos significativo (LSB).

Esta estrutura sugere a existência de correlações não-lineares e localizadas que uma análise puramente sequencial (byte-a-byte) poderá não capturar eficazmente.

## 2.2 Limites Teóricos e Análise Global

O limite teórico fundamental para a compressão sem perdas é dado pela **Entropia de Shannon**. Considerando o ficheiro como uma fonte de memória nula $X$ que gera símbolos $x \in \{0, \ldots, 255\}$, a entropia de ordem-0 é definida por:

$$H(X) = -\sum_{i=0}^{255} P(x_i) \log_2 P(x_i) \quad [\text{bits/símbolo}] \tag{1}$$

Experimentalmente, ao aplicar a Equação 1 à totalidade do *payload* binário, obteve-se:

$$H(X) \approx 6.22 \textbf{ bits/byte}$$

Este valor indica que, ignorando qualquer dependência entre bytes, a compressão máxima teórica seria de apenas $\sim 22\%$. Para investigar dependências sequenciais, calculou-se a **Entropia Condicional** de primeira ordem, que mede a incerteza de um símbolo $X_n$ dado o conhecimento do anterior $X_{n-1}$:

$$H(X|Y) = -\sum_{y \in \mathcal{X}} P(y) \sum_{x \in \mathcal{X}} P(x|y) \log_2 P(x|y) \tag{2}$$

O resultado experimental obtido foi:

$$H(X_n|X_{n-1}) \approx 5.36 \textbf{ bits/byte}$$

Embora $H(X|Y) < H(X)$, confirmando a existência de correlação inter-simbólica (teorema do condicionamento reduz a entropia), o valor 5.36 permanece elevado. A nossa hipótese é que a natureza intercalada dos dados BF16 (MSB estruturado seguido de LSB ruidoso) "mascara" a verdadeira correlação entre os pesos adjacentes.

## 2.3 Análise Estrutural Diferenciada (*Byte-Splitting*)

Para validar a hipótese de que a entropia está concentrada no byte da mantissa, procedeu-se à separação do fluxo de dados em dois canais distintos: *Stream MSB* (bytes ímpares) e *Stream LSB* (bytes pares).

As entropias de ordem-0 foram recalculadas individualmente para cada canal:

Tabela 1: Comparação de Entropia por Canal (Split)

| Canal | Conteúdo | Entropia Medida ($H$) | Característica |
|---|---|---|---|
| **MSB** | Expoente/Sinal | **2.71 bits/byte** | Altamente Estruturado |
| **LSB** | Mantissa | **7.96 bits/byte** | Ruído Quase Uniforme |

### Evidência Visual

Os histogramas de frequência (Figuras 1 e 2) corroboram os valores numéricos.

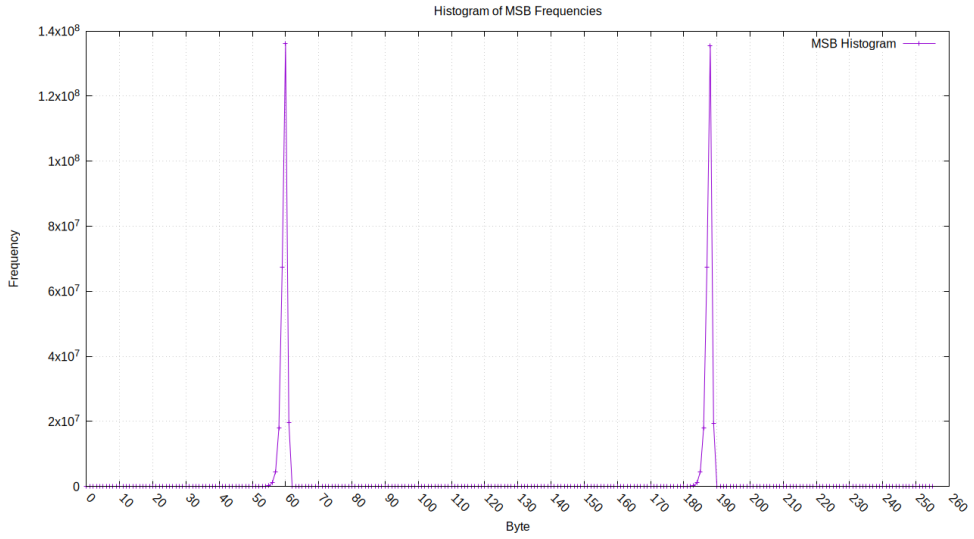


Figura 1: Histograma do Byte Mais Significativo (MSB). Nota-se uma distribuição Laplaciana acentuada, típica de pesos de redes neuronais normalizados, justificando o valor baixo de $H \approx 2.71$.

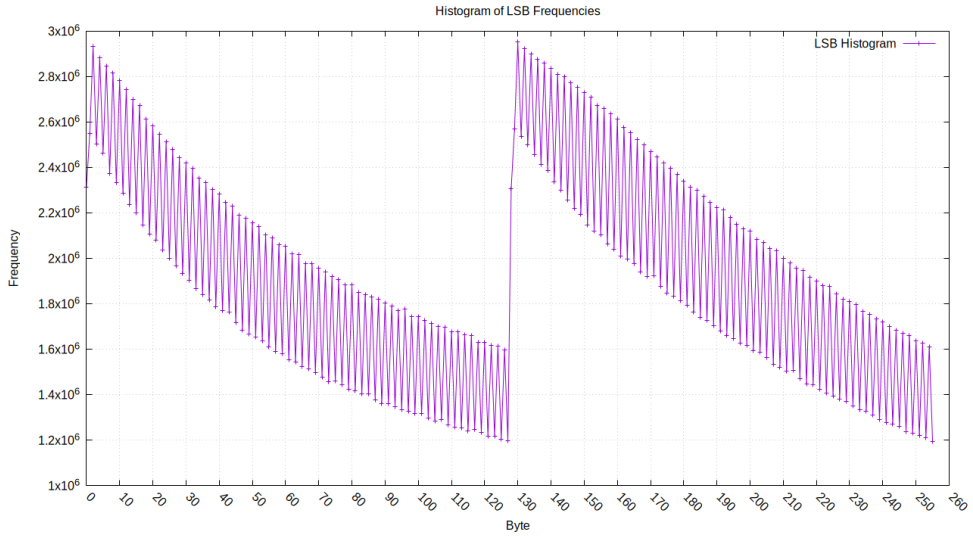


Figura 2: Histograma do Byte Menos Significativo (LSB). A distribuição aproxima-se da uniforme (plana), o que explica a entropia de $H \approx 7.96$, muito próxima do máximo teórico de 8 bits.

### 2.4 Síntese e Estratégia Adotada

A média das entropias separadas é $(2.71 + 7.96)/2 \approx 5.34$, um valor virtualmente idêntico à Entropia Condicional global (5.36). Isto leva-nos a concluir que a "memória" da fonte detetada na análise global era, na verdade, a estrutura interna do formato BF16 e não apenas correlação sequencial.

**Nota**

Com base nestes dados teóricos e experimentais, a estratégia de compressão será:

1. **Pré-processamento (Split):** Separar os fluxos para isolar o ruído da estrutura.
2. **Canal LSB:** Dado que $H \approx 8$, não justifica custo computacional elevado. Será armazenado com compressão mínima ou nula.
3. **Canal MSB:** Dado que $H \approx 2.71$, este canal é o candidato ideal para **Codificação Preditiva** (Delta) seguida de **Codificação Entrópica** (Huffman ou Aritmética), visando reduzir a entropia residual para valores próximos de 2 bits/byte.

## 3 Implementação e Otimização do Canal MSB

Com base na análise preliminar, o canal MSB (Byte Mais Significativo) apresentava uma entropia de base de 2.70 bits/byte, tornando-se o foco principal para ganhos de compressão. Esta secção descreve o processo iterativo de otimização, desde as tentativas de modelação preditiva até à seleção do codificador de entropia final.

### 3.1 Avaliação de Técnicas de Transformação Preditiva

A literatura de compressão de dados (e.g., JPEG Lossless, Audio Coding) sugere frequentemente o uso de codificação preditiva para reduzir a variância dos resíduos em sinais correlacionados. Testaram-se duas abordagens para explorar a correlação sequencial entre os pesos do modelo.

### 3.1.1 Abordagem 1: Preditor Linear Aritmético (Delta)

A primeira tentativa utilizou um preditor de primeira ordem clássico, onde o resíduo $r_n$ é calculado pela diferença aritmética entre o byte atual $x_n$ e o anterior $x_{n-1}$:

$$r_n = (x_n - x_{n-1}) \mod 256 \tag{3}$$

**Resultado:** Contrariamente ao esperado, a aplicação deste preditor **aumentou** a entropia do canal MSB de 2.70 para 3.28 bits/byte (um ganho negativo de $-0.58$ bits).

**Análise de Falha:** A inspeção dos dados revelou que este comportamento se deve ao formato `BF16`. O bit mais significativo do MSB corresponde ao *Sinal*. Quando os pesos da rede neuronal oscilam entre valores pequenos positivos e negativos (comum em LLMs), o bit de sinal alterna (e.g., de 0 para 1), o que a subtração aritmética interpreta como um "salto" numérico de grande magnitude (e.g., +128), dispersando o histograma dos resíduos.

### 3.1.2 Abordagem 2: Preditor Lógico (XOR)

Para mitigar o problema do bit de sinal, implementou-se um preditor baseado na operação exclusiva-ou (XOR), comum em compressores de ponto flutuante como o FPC, para capturar semelhanças de padrões de bits independentemente do valor aritmético:

$$r_n = x_n \oplus x_{n-1} \tag{4}$$

**Resultado:** Embora tenha apresentado um desempenho superior ao preditor aritmético, a entropia resultante foi de 3.11 bits/byte, ainda superior à entropia original de 2.70 bits.

### 3.1.3 Decisão de Engenharia

Concluiu-se que a baixa entropia do canal MSB (2.70 bits) não advém da correlação sequencial imediata ($x_n \approx x_{n-1}$), mas sim da distribuição global dos expoentes (distribuição bimodal estática). Qualquer tentativa de transformação preditiva simples tende a destruir esta estrutura estatística favorável. Consequentemente, optou-se por **não aplicar transformações** e codificar diretamente os valores brutos do canal MSB.

## 3.2 Codificação de Entropia: Huffman vs. Aritmética

Após definir que os dados seriam codificados sem transformação prévia, comparou-se o desempenho de dois algoritmos de entropia para o passo final de compressão.

Tabela 2: Comparação de Desempenho no Ficheiro Completo (Canal MSB)

| **Algoritmo** | **Tamanho Final** | **Tempo Total** | **Rácio Efetivo** |
|---|---|---|---|
| Huffman (Estático) | 633.04 MB | **13.98 s** | 1.58:1 |
| Aritmético (Estático) | **631.00 MB** | 19.07 s | 1.59:1 |

A Codificação Aritmética permitiu uma redução adicional de $\approx 2$ MB. Teoricamente, isto deve-se à sua capacidade de alocar um número fracionário de bits por símbolo, aproximando-se do limite da entropia (2.70), enquanto o Huffman é penalizado pela restrição de usar números inteiros de bits. Contudo, este ganho marginal (0.3%) implicou um custo computacional de 36% no tempo de execução.

### 3.3 Definição dos Modos de Operação

Para cumprir os requisitos de projeto que exigem diferentes pontos de operação (compromisso tempo/compressão), definiram-se dois modos no compressor final:

- **Modo *Fast*:** Utiliza o algoritmo de **Huffman**. Prioriza a velocidade de processamento, ideal para cenários de inferência em tempo real ou carregamento rápido de modelos.
- **Modo *Best*:** Utiliza a **Codificação Aritmética**. Prioriza a minimização absoluta do espaço em disco, ideal para arquivo ou distribuição em redes com largura de banda limitada.

## 4 Conclusões