Modelo de Treinamento Completo para RNA de Conversa

Este documento define o MODELO DE TREINO IDEAL para uma RNA de conversa híbrida (local + LLM/Ollama), focada em diálogo natural, contexto, memória e aprendizado incremental.

1. Objetivo do Modelo de Treino

O objetivo é permitir que a RNA aprenda conversação de forma progressiva, estruturada e expansível, sem depender exclusivamente de um único modelo.

2. Princípios Fundamentais

- Treinar intenção, não apenas resposta.
- Manter memória de curto e longo prazo.
- Permitir correção explícita do usuário.
- Funcionar com ou sem LLM externo.

3. Camada 1 – Normalização de Entrada

Todo texto deve ser normalizado antes do treino ou inferência, reduzindo ruído linguístico.

Formato de treino: ENTRADA_USUARIO texto_original texto_normalizado tokens_chave

4. Camada 2 – Intenções (Intents)

Define o objetivo do usuário. Cada intenção deve conter dezenas de variações.

Formato: INTENT nome exemplos respostas_base

5. Camada 3 – Pergunta e Resposta (Q&A;)

Treino supervisionado direto, usado principalmente em fallback e respostas factuais.

Formato: QA pergunta resposta tags fonte

6. Camada 4 – Diálogos com Contexto

Ensina fluxo real de conversa, coerência e sequência lógica.

Formato: DIALOGO contexto turnos (usuario/bot)

7. Camada 5 – Memória e Estado

Define o que a RNA deve lembrar sobre o usuário ou a conversa.

Formato: MEMORIA tipo chave valor

8. Camada 6 – Correções (Aprendizado Supervisionado Real)

Permite aprendizado real a partir de erros corrigidos pelo usuário.

Formato: CORRECAO entrada_usuario resposta_bot correcao_usuario resposta_correta

9. Camada 7 – Fallback Inteligente

Define comportamento quando a RNA não tem confiança suficiente.

Formato: FALLBACK condicoes respostas

10. Camada 8 – Integração com LLM (Ollama)

Controla quando e como delegar respostas a um LLM local.

Formato: LLM_POLICY quando_usar prompt_base

11. Camada 9 – RAG Local (Recuperação de Conhecimento)

Ensina a RNA a buscar informações relevantes antes de responder.

Formato: RAG fonte top_k limiar

12. Conclusão

Este modelo de treino permite construir uma RNA de conversa robusta, evolutiva e profissional, adequada tanto para uso local quanto híbrido com modelos LLM como Ollama.