

Resumo executivo

O arquivo anonimizado apresenta erros graves de substituição que comprometem a legibilidade e a confiabilidade dos dados. As principais causas são limites de token inadequados e regras de mapeamento muito amplas, que trocaram palavras de uso comum pelo dicionário fictício. O relatório abaixo descreve cada falha, aponta prováveis origens no pipeline (spaCy + regex + Faker) e oferece sugestões técnicas — da criação de regras no EntityRuler à adoção de placeholders categóricos — para que o desenvolvedor corrija o fluxo.

1. Problemas identificados

1. Termos funcionais substituídos (ex.: “Horário” → “Azevedo”) — Perda de sentido jurídico do documento.
2. Concatenação ou truncagem de tokens (“Almeidaimadamente”) — Texto ilegível e risco de quebras em parsers automáticos.
3. Dados fictícios mantêm formato real (CPF, telefone) — Ainda parecem dados válidos, ferindo princípios de minimização da LGPD.
4. Vocabulário inconsistente (nomes não realistas para cargos/locais) — Diminui a credibilidade da anonimização.

2. Causas prováveis

1. Boundary de tokens não respeitado — Substituições feitas por regex sem \b ou sem consulta ao tipo de entidade, permitindo matches parciais em palavras comuns.
2. Ausência de dicionário de exclusões (stop■terms) — Palavras de domínio (“Delegacia”, “Horário”) não foram bloqueadas antes da troca.
3. Modelo NER genérico — SpaCy PT apresenta falsos positivos em textos jurídicos sem fine■tuning.
4. Faker com formatação realista — O provider pt_BR gera CPFs verossímeis por design.
5. Pipeline linear — Falta validação pós■substituição para capturar distorções antes de salvar a nova versão.

3. Recomendações técnicas

3.1 Ajustar a detecção de entidades

- * Treinar NER customizado com amostras de oitivas (transfer learning + `spacy train`) para reduzir falsos positivos.
- * Adicionar `EntityRuler` com padrões específicos (regex de CPF, telefone, nomes iniciando por maiúscula e não presentes em lista branca).
- * Criar lista branca / stop■terms para termos que jamais devem ser substituídos (Delegacia, Juiz, Horário).

3.2 Estratégia de substituição

- * Trocar valores sensíveis por placeholders semiformatados (`[CPF]`, `[TEL]`, `[NOME_1]`) em vez de dados sintéticos completos.
- * Se dados sintéticos forem obrigatórios, gerar CPFs inválidos (dígito verificador errado) ou marcar explicitamente como fictícios.

3.3 Pós■validação

- * Rodar diff automatizado garantindo que nenhum termo da lista branca foi alterado.
- * Usar ferramenta externa (ex.: scrubadub) como auditoria de PII residual.

3.4 Robustez do código

- * Regex com delimitadores de palavra (``\bCPF\b``) para evitar matches parciais.
- * Desabilitar pipes desnecessários durante substituição para preservar offsets.
- * Versionar e empacotar regras (``patterns.json``) via ``spacy package``.

4. Próximos passos sugeridos

1. Refatorar pipeline aplicando as recomendações acima.
2. Criar conjunto de testes com pelo menos 30 oitivas marcadas e oráculo de saída; meta de precisão $\geq 0,95$.
3. Documentar versões e parâmetros (spaCy, Faker seed) para reprodutibilidade.
4. Conduzir revisão manual dos primeiros 10 documentos após refatoração antes de produção.

5. Referências

- * Custom NER com spaCy em português.
- * Documentação do EntityRuler.
- * Documentação Faker pt_BR.
- * Artigos sobre boas práticas de anonimização e LGPD.