IA GENERATIVA E DIREITOS AUTORAIS – PL 2.338/23

Rony Vainzof – rony@vlklaw.com.br

FECOMERCIOSP | sesc senac

Rony Vainzof (Consultor FecomercioSP) – rony@vlklaw.com.br

Agenda

1) Equilíbrio e efeitos colaterais
2) Como funciona a tecnologia
3) Fair Training
4) Direito Comparado
5) Sugestão de texto

Equilíbrio

Criatividade e conteúdo humano precisam ser preservados, recompensados e remunerados

Regras rígidas e limitantes de direitos autorais para o treinamento da IAG podem trazer efeitos colaterais preocupantes, como:

- Custos proibitivos para startups e empresas de pequeno porte, aumentando a vantagem competitiva das big techs;
- Fuga de centros de IA para países mais permissivos;
- Menor precisão diante da menor quantidade de dados; e
- Sufocar a pesquisa aberta e concentrar inovação em ambientes fechados.

Tecnologia não se importa com o conteúdo enquanto obra

- O conteúdo enquanto obra protegível é utilizado somente como **insumo técnico** para ensinar o modelo sobre as relações estatísticas entre os seus elementos.
- Há a fragmentação do conteúdo coletado: **tokens + vetores matemáticos**
- Por exemplo, “rei” e “rainha” são **tokenizados** e ficam próximos nesse espaço (vetores matemáticos), enquanto “rei” e “banana” se distanciam.
- **Modelo somente aprende padrões estatísticos** (de linguagem, no caso de LLM).
- Modelo **não guarda** cada obra de forma individual, mas extrai padrões estatísticos gerais a partir do conjunto de informações coletadas em massa e tokenizadas.
- A **memorização** de trechos específicos pode ocorrer, mas em **pequena escala** e, em geral, restrita a conteúdos raros.
- **Modelo generaliza** e o **impacto de cada obra isolada se dilui** dentro da massa de dados.
- **Não** há como **rastrear a contribuição unitária**, nem estruturar um sistema de metadados que permita vincular entradas (inputs) a saídas (outputs).
- É **inadequado** tratar o treinamento desses **modelos como equivalente ao uso individualizado de uma obra** musical, jornalística ou literária.

How much do language models memorize?

John X. Morris[1,3], Chawin Sitawarin[2], Chuan Guo[1], Narine Kokhlikyan[1], G. Edward Suh[3,4], Alexander M. Rush[3], Kamalika Chaudhuri[1], Saeed Mahloujifar[1]

[1]FAIR at Meta, [2]Google DeepMind, [3]Cornell University, [4]NVIDIA

Date: June 19, 2025

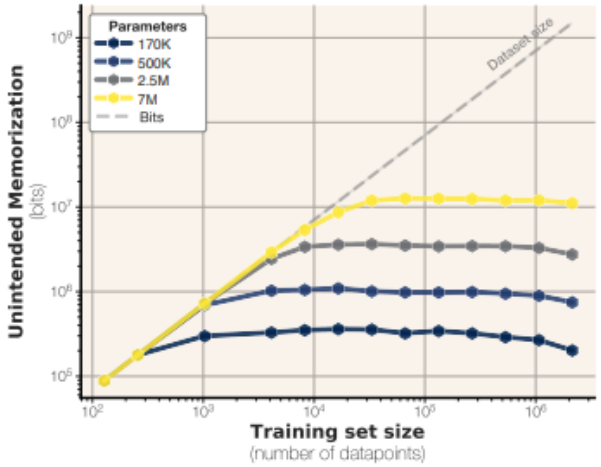

Figure 1 Unintended memorization of uniform random data (Section 3). Memorization plateaus at the empirical capacity limit of different-sized models from the GPT-family, approximately 3.6 bits-per-parameter.

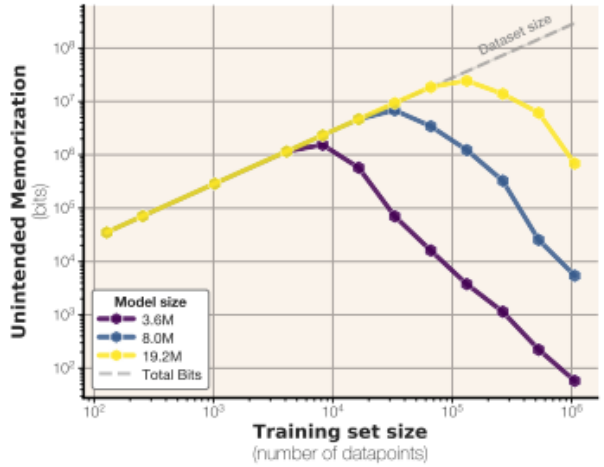

Figure 2 Unintended memorization of text across model and dataset sizes (Section 4). All quantities are calculated with respect to a large oracle model trained on the full data distribution.

https://arxiv.org/pdf/2505.24832

Quando estudamos algo, podemos **decorar frases exatas** (memorização) ou **entender o assunto e aplicar em outras situações** (generalização).

1.Todo modelo tem um limite de memória. Nos modelos do tipo GPT, cada "pedacinho" do modelo consegue guardar cerca de **3,6 bits de informação**. Ou seja, é possível armazenar alguns fragmentos exatos, **mas não livros inteiros.**

2.No começo do treino, o modelo guarda muita informação ao pé da letra (memorização).

3.Depois de um ponto, se o conjunto de dados fica muito grande, o modelo "entende o padrão" e passa a **generalizar**, em vez de repetir.

4.Essa virada é o **"grokking"** - quando o modelo para de decorar e começa a realmente aprender padrões.

Fair training

(Doutrina do Uso Justo para o Treinamento de IA)

1) Juiz do Distrito Norte da Califórnia, nos EUA (Anthropic - Claude LLM): "todo mundo também lê textos e depois escreve novos textos. Obrigar alguém a pagar especificamente pelo uso de um livro cada vez que o lê, cada vez que o recorda de memória, cada vez que posteriormente o consulta para escrever coisas novas de novas maneiras seria impensável". No entanto, considerou ilícita a prática de baixar e manter livros pirateados para construir uma biblioteca digital permanente.

2) Brasil: LDA + STJ: se tratar de situação especial; não prejudicar a exploração normal da obra; e não causar dano injustificado aos interesses do autor.

3) Argumentos favoráveis:
    1) Os dados são utilizados apenas como insumos técnicos para ensinar padrões estatísticos e não para copiar as obras originais;
    2) O aprendizado de máquina é comparável ao processo humano de indução e generalização; e
    3) **A responsabilização continua possível em relação aos outputs que eventualmente violem direitos autorais.**

Direito Comparado

1) **União Europeia:** permite a mineração de texto e dados para qualquer outra finalidade, contanto que os titulares de direitos autorais não tenham reservado expressamente seus direitos (opt-out). O EU AI Act exige transparência (resumo de datasets) e política de respeito a direitos autorais.

2) **Japão**: é permitido a utilização de obras protegidas por direitos autorais para fins de "análise de informações" sem a necessidade de permissão do detentor dos direitos.

3) **EUA**: como visto, há tendência pelo uso justo, mas o seu Escritório de Direitos Autorais aponta teses favoráveis e contrárias.

4) **Brasil**: o atual texto do PL 2.338/23 é restritivo.

Premissas defendidas para o PL 2.338/23

1) Treinamento de GPAI não se importa com o conteúdo enquanto obra protegida (token, vetores e padrões matemáticos estatísticos);

2) Não há, em regra, armazenamento do conteúdo protegido;

3) Não há como rastrear a contribuição unitária nem estruturar um sistema de metadados que permita vincular entradas (inputs) a saídas (outputs);

4) É inadequado tratar o treinamento desses modelos como se fosse equivalente ao uso individualizado de uma obra musical, jornalística ou literária;

5) O controle de violação de direitos autorais deve ser feito no output;

6) Permitir o treinamento de IA a partir do uso de dados publicamente disponíveis na internet, desde que sem opt-out;

7) Criação de sistema opt-out: com o opt-out, o titular de direitos de autor poderá negociar diretamente o licenciamento oneroso das obras sob sua titularidade com os desenvolvedores à sua escolha ou com bancos de dados;

8) Desenvolvedor de IA que utilizar conteúdo protegido por direitos autorais deve publicar sumários informando os tipos de materiais utilizados.

REGULAÇÃO NÃO É INIMIGA DA INOVAÇÃO...

MAS A MÁ REGULAÇÃO SIM!

Rony Vainzof (Consultor FecomercioSP) – rony@vlklaw.com.br