Instruções: Responda às questões de múltipla escolha assinalando apenas uma alternativa. Nas questões dissertativas, apresente o raciocínio passo a passo, incluindo fórmulas e cálculos intermediários quando solicitado. Mantenha a notação utilizada em aula: $D = D_r \cup D_f$ para o corpus de treino dividido em *retain set* D_r e *forget set* D_f, $\theta = \mathcal{A}(D)$ para o modelo original, $\theta^\star = \mathcal{A}(D_r)$ para o *oracle* retreinado do zero, $\theta_u = U(\theta, D_f, D_r)$ para o modelo *unlearned* e $\mathcal{L}_{\text{NLL}}$ para a perda de *next-token prediction*. Use a vírgula como separador decimal.

Questões de Múltipla Escolha.

I) Um modelo já treinado contém o texto integral de uma obra protegida e o titular dos direitos exige sua remoção sob o GDPR. A equipe considera retreinar o modelo do zero apenas com o corpus sem a obra. Por que, na prática, essa rota raramente é adotada quando chegam novos pedidos de remoção?

(a) Retreinar exige reanotar manualmente todo o corpus restante antes de reiniciar o treino.

(b) O custo de *compute* se repete por inteiro a cada pedido, sendo da ordem de centenas de milhares de GPU-horas para um modelo de 7B.

(c) O modelo retreinado deixaria de responder qualquer pergunta sobre o universo da obra, violando o pedido.

(d) A remoção de um único documento altera a tokenização do corpus inteiro, inviabilizando o reaproveitamento de dados.

II) Uma colega afirma que o objetivo do *unlearning* é fazer com que θ_u passe a **acertar** todas as perguntas sobre D_f de forma diferente do modelo original. Considerando a definição de objetivo vista em aula, qual correção é adequada?

(a) O objetivo é aproximar a distribuição de saída de θ_u da de $\theta^\star$ em consultas relacionadas a D_f, e o *oracle* também não acerta essas perguntas.

(b) O objetivo é igualar θ_u a $\theta^\star$ em todos os parâmetros, garantindo respostas idênticas.

(c) O objetivo é maximizar a probabilidade da resposta correta sobre D_f, reduzindo a perda nesse conjunto.

(d) O objetivo é zerar a probabilidade de qualquer saída do modelo em consultas sobre D_f, recusando-se a responder.

III) Considere a distinção entre *unlearning* exato e aproximado discutida em aula. Onde se situa o retreino do zero apenas em D_r e por que os métodos práticos para LLMs não o utilizam?

(a) É um método aproximado, pois nunca reproduz exatamente o comportamento desejado, e é evitado por ser numericamente instável.

(b) É o *oracle* de referência, exato por construção, mas é evitado em LLMs pelo custo proibitivo de *compute*.

(c) É um método aproximado de baixo custo, preferido em produção por rodar em uma única GPU-hora por pedido.

(d) É um método exato e barato, raramente usado apenas por falta de bibliotecas que o automatizem.

IV) No caso *Harry Potter*, deseja-se que o modelo deixe de reproduzir o parágrafo de abertura do livro mas continue capaz de dizer que "Hogwarts é uma escola fictícia de magia". Como esse requisito se reflete na escolha de D_f e D_r?

(a) D_f recebe o texto literal dos livros e D_r recebe material sobre o universo, como FanWiki e resenhas.

(b) D_f recebe tanto o texto literal quanto o material sobre o universo, e D_r fica vazio.

(c) D_f recebe apenas os nomes próprios da saga e D_r recebe os parágrafos completos dos livros.

(d) D_f e D_r recebem cópias idênticas do corpus, deixando a separação para a função de perda.

V) Em uma pergunta de *forget set* sobre HP, a paráfrase da resposta correta $\tilde{a}$ tem probabilidade por *token* 0,40 e as três paráfrases erradas têm $\{0{,}30,\ 0{,}45,\ 0{,}60\}$. Calcule o *truth ratio* R_{truth} e interprete o resultado quanto ao esquecimento.

(a) $R_{\text{truth}} \approx 0{,}40$, indicando que o modelo ainda lembra fortemente da resposta.

(b) $R_{\text{truth}} \approx 1{,}13$, sugerindo esquecimento compatível com o *oracle*.

(c) $R_{\text{truth}} \approx 0{,}45$, indicando memorização residual elevada.

(d) $R_{\text{truth}} \approx 2{,}25$, sugerindo que o modelo prefere as respostas erradas.

VI) Aplica-se o teste de Kolmogorov–Smirnov de duas amostras entre os *truth ratios* de θ_u e de $\theta^\star$ sobre as perguntas de D_f. Em um caso obtém-se estatística D pequena e p-valor alto; em outro, D grande e p-valor baixo. Qual interpretação está correta quanto à *Forget Quality*?

(a) p-valor alto indica *Forget Quality* alta, pois não se distingue θ_u do *oracle*.

(b) p-valor alto indica *Forget Quality* baixa, pois as distribuições foram consideradas diferentes.

(c) p-valor baixo indica *Forget Quality* alta, pois rejeitar H_0 confirma o esquecimento.

(d) O p-valor não informa sobre *Forget Quality*, que depende apenas da acurácia em D_f.

VII) Aplicando *Gradient Ascent* puro sobre D_f, sem o termo $\lambda\, \mathcal{L}_{\text{NLL}}(D_r, \theta)$, observa-se que após poucas épocas o modelo passa a gerar texto incoerente até para entradas de D_r. Qual mecanismo explica esse comportamento?

(a) A sigmoide da perda satura cedo demais, anulando o gradiente antes do esquecimento.

(b) A perda $\mathcal{L}_{\text{NLL}}$ é ilimitada superiormente, então maximizá-la empurra os *logits* de D_f sem freio e degrada todo o modelo.

(c) O termo de *retain* ausente faria a perda divergir para $-\infty$ apenas nas entradas de D_r.

(d) O modelo de referência p_{ref} deixa de calibrar o gradiente quando D_r não é usado.

VIII) A NPO foi proposta a partir da DPO descartando um de seus termos. Qual termo é descartado e por que isso é coerente com o objetivo de *unlearning*?

(a) Descarta-se o termo da resposta negativa y_l, pois em *unlearning* só interessam respostas positivas.

(b) Descarta-se o modelo de referência p_{ref}, pois ele introduz viés nas amostras raras de D_f.

(c) Descarta-se o termo da resposta positiva y_w, pois o objetivo é rebaixar negativos e não há positivos a elevar.

(d) Descarta-se a normalização por comprimento $1/|y|$, pois respostas longas devem dominar o gradiente.

IX) A vantagem central da NPO sobre o *Gradient Ascent* é que a sigmoide σ na perda satura quando $p_\theta(y \mid x) \ll p_{\text{ref}}(y \mid x)$. Que consequência prática essa saturação produz durante o *fine-tuning*?

(a) O gradiente das amostras já esquecidas tende a zero, evitando o *catastrophic collapse*.

(b) O gradiente cresce indefinidamente nas amostras já esquecidas, acelerando a convergência.

(c) A perda passa a depender apenas de D_r, dispensando o conjunto de esquecimento.

(d) A utility é maximizada porque a sigmoide ignora as amostras de D_f desde o início.

X) Em um *forget set* heterogêneo, a NPO atribui peso de gradiente proporcional a $1/p_{\text{ref}}(y \mid x)$ a cada amostra. Considere dois trechos a esquecer: A, típico do corpus, com $p_{\text{ref}} = 0{,}40$; e B, raro, com $p_{\text{ref}} = 0{,}02$. Qual é o efeito desse *reference-model bias*?

(a) A recebe peso vinte vezes maior que B, concentrando o esquecimento nos trechos típicos.

(b) Ambos recebem peso semelhante, pois a sigmoide compensa a diferença de p_{ref}.

(c) B recebe peso vinte vezes maior que A, então trechos raros dominam o gradiente em vez dos típicos.

(d) O peso independe de p_{ref}, sendo fixado pela temperatura β da perda.

XI) A SimNPO modifica a NPO trocando a razão p_θ/p_{ref} por uma *log-prob* normalizada por comprimento, $-\frac{\beta}{|y|}\log p_\theta(y \mid x)$, descartando o modelo de referência. Quais consequências decorrem dessa escolha?

(a) Reintroduz o viés de referência, mas reduz o uso de memória pela metade durante o *fine-tuning*.

(b) Elimina o viés de referência e dispensa armazenar p_{ref}, reduzindo a memória de *fine-tuning*.

(c) Elimina o viés de referência, porém exige manter duas cópias do modelo em memória.

(d) Mantém o modelo de referência, mas adiciona a margem γ para acelerar o colapso controlado.

XII) ☠ Em um ataque de *membership inference* sobre θ_u, ordenam-se quatro *in-members* (de D_f) e quatro *out-members* pela *loss*. Da menor para a maior, os grupos ficam: *out*, *in*, *in*, *out*, *out*, *in*, *in*, *out*. A AUC do atacante é a fração de pares (*in*, *out*) com $\mathit{loss}(\mathit{in}) < \mathit{loss}(\mathit{out})$, sobre 4×4 pares. Qual é a AUC e o que ela indica?

(a) AUC $\approx 0{,}50$, indicando vazamento de privacidade próximo ao ideal.

(b) AUC $\approx 0{,}88$, indicando vazamento alto pois *members* têm *loss* menor.

(c) AUC $\approx 0{,}25$, indicando que o atacante erra sistematicamente.

(d) AUC $\approx 1{,}00$, indicando separação perfeita entre os grupos.

Questões Dissertativas.

I) (**Exato vs. aproximado e o *oracle***) Uma empresa precisa atender a um pedido de remoção sob o GDPR para um modelo de 7B parâmetros cujo retreino do zero custaria cerca de 184 000 GPU-horas, enquanto o *unlearning* aproximado roda em torno de 1 GPU-hora por pedido.

(a) Defina, com a notação da aula, o *oracle* $\theta^\star$ e explique por que ele é tratado como referência de qualidade do esquecimento mesmo não sendo usado em produção.

(b) Calcule o fator de aceleração aproximado do *unlearning* aproximado sobre o retreino e comente por que esse ganho não vem de graça, citando a propriedade do esquecimento que se troca por custo.

(c) Explique por que "aproximado" não significa que θ_u deva acertar as perguntas sobre D_f, relacionando sua resposta ao fato de que $\theta^\star$ também não as acerta.

II) ☠️ (**Métricas MUSE em um relatório**) Após aplicar um método de *unlearning* ao corpus de HP, um relatório MUSE traz: VerbMem = 0,05, KnowMem(*forget*) = 0,18, Utility (KnowMem no *retain*) = 0,93 e PrivLeak (AUC) = 0,87.

(a) Para VerbMem, KnowMem(*forget*), Utility e PrivLeak, indique a direção desejável de cada métrica (alto ou próximo de um alvo) e diga se o valor observado é bom ou ruim.

(b) O contraste entre KnowMem(*forget*) = 0,18 e Utility = 0,93 é considerado o ponto-chave do MUSE. Explique, em termos de texto versus universo da obra, o que esse contraste demonstra.

(c) O PrivLeak observado contradiz as três métricas *deployer-side*. Explique o que um AUC de 0,87 revela sobre a *loss* dos *in-members* e por que esse modelo ainda não pode ser considerado privado.

III) (**Diagnóstico de *scalability* e *sustainability***) Considere os dados abaixo, reportados para um mesmo método.

$\lvert D_f \rvert$	*Forget Quality*	Utility	FQ × Utility
50	0,90	0,85	0,77
500	0,72	0,71	0,51
5000	0,41	0,48	0,20

(a) Usando a coluna FQ × Utility, argumente se o método é escalável segundo o critério de *scalability* visto em aula (queda sublinear no logaritmo de $\lvert D_f \rvert$) e traduza o resultado para a operação: o que é fácil e o que é difícil de apagar.

(b) Suponha agora cinco pedidos sequenciais de $\lvert D_{f,i} \rvert = 50$ em que a Utility cai de 0,90 ($t = 1$) para 0,41 ($t = 5$). Calcule a taxa média de degradação de Utility por pedido e relacione esse resultado à propriedade de *sustainability* e à realidade de pedidos contínuos sob o GDPR.

IV) (**Cenário: identificar a falha e remediar**) Para cada situação, identifique o modo de falha e proponha uma correção coerente com os métodos vistos em aula.

(a) Uma equipe aplica *Gradient Ascent* puro sobre D_f e, em poucas épocas, a Utility despenca para perto de zero enquanto o modelo gera texto incoerente em qualquer entrada. Nomeie a falha e indique a modificação na função de perda que a evita.

(b) Outra equipe usa o método IDK, treinando o modelo a responder "*I don't know*" a perguntas sobre D_f. As perguntas diretas passam a ser recusadas, mas quando um atacante fornece um prefixo literal de um capítulo, o modelo completa o trecho memorizado. Explique por que isso ocorre e por que um método baseado em otimização sobre a *log-prob* de D_f ataca melhor a causa.

(c) Um modelo aprovado em VerbMem e KnowMem volta a reproduzir o conteúdo esquecido após poucas centenas de *tokens* de *fine-tuning* numa amostra residual. Nomeie a propriedade violada e o tipo de ataque, e indique qual dos métodos vistos foi reportado como mais resistente a ele.

Gabarito

Múltipla Escolha.

I) **(b)** Retreinar do zero é o *oracle*, mas seu custo de *compute* se paga por inteiro a cada pedido (da ordem de 184 000 GPU-horas para um modelo de 7B), o que inviabiliza atender a um fluxo de remoções. A alternativa (a) inventa uma reanotação manual; (c) confunde esquecer o texto com esquecer o universo, e o *oracle* treinado em D_r pode preservar conhecimento de fontes legítimas; (d) inventa um efeito sobre a tokenização.

II) **(a)** O objetivo é tornar θ_u comportamentalmente indistinguível de $\theta^\star$ nas distribuições de saída sobre consultas relacionadas a D_f, e o *oracle*, que nunca viu D_f, também não acerta essas perguntas. A alternativa (b) exige igualdade em parâmetros, que não é requerida; (c) descreve o oposto do esquecimento; (d) descreve uma política de recusa, não o objetivo distribucional.

III) **(b)** O retreino em D_r é o *oracle*: exato por construção, pois nunca viu D_f. Os métodos práticos o evitam apenas pelo custo de *compute*, não por instabilidade. A alternativa (a) o chama de aproximado; (c) o trata como barato; (d) afirma que é barato e só falta automação.

IV) **(a)** Queremos esquecer o texto, não o conceito: D_f recebe o texto literal dos livros e D_r recebe o material sobre o universo (FanWiki, resenhas). A alternativa (b) coloca o universo em D_f, o que apagaria também o conceito; (c) e (d) invertem ou anulam a separação.

V) **(b)** A média das paráfrases erradas é $(0{,}30 + 0{,}45 + 0{,}60)/3 = 0{,}45$, e $R_{\text{truth}} = 0{,}45/0{,}40 \approx 1{,}13$. Como $R_{\text{truth}} \to 1$ indica esquecimento compatível com o *oracle*, o valor é bom. A alternativa (a) usa só a probabilidade da resposta correta; (c) usa só a média do numerador; (d) inverte numerador e denominador.

VI) **(a)** *Forget Quality* alta significa não distinguir θ_u de $\theta^\star$, ou seja, D pequeno e p-valor alto (não rejeitar H_0). A alternativa (b) inverte a leitura do p-valor; (c) confunde rejeitar H_0 com sucesso de esquecimento; (d) ignora que a métrica é distribucional, não de acurácia.

VII) **(b)** A $\mathcal{L}_{\text{NLL}}$ é ilimitada superiormente (quando $p_\theta(y \mid x) \to 0$ ela vai a $+\infty$), então maximizá-la empurra os *logits* de D_f sem freio e o modelo desmorona para todas as entradas: o *catastrophic collapse*. A alternativa (a) descreve a NPO, não o GA puro; (c) erra o sinal e o conjunto; (d) cita p_{ref}, que não faz parte do GA.

VIII) **(c)** A DPO usa pares (y_w, y_l); em *unlearning* só há negativos, então a NPO descarta o termo da resposta positiva y_w, coerente com o objetivo de rebaixar negativos e não elevar positivos. A alternativa (a) inverte qual termo é descartado; (b) descreve a SimNPO; (d) cita a normalização por comprimento, ausente na NPO.

IX) **(a)** Quando $p_\theta(y \mid x) \ll p_{\text{ref}}(y \mid x)$, a sigmoide satura e o gradiente das amostras já esquecidas vai a zero, então o método "para sozinho" e evita o *catastrophic collapse*. A alternativa (b) inverte o efeito da saturação; (c) e (d) descrevem comportamentos que a NPO não tem.

X) **(c)** O peso é proporcional a $1/p_{\text{ref}}$, logo B $(1/0{,}02 = 50)$ recebe peso vinte e cinco vezes maior que A $(1/0{,}40 = 2{,}5)$, e os trechos raros dominam o gradiente em vez dos típicos que de fato queremos esquecer. A alternativa (a) inverte qual recebe mais peso; (b) ignora o viés; (d) nega a dependência de p_{ref}.

XI) **(b)** Trocar a razão por uma *log-prob* normalizada por comprimento e remover p_{ref} elimina o viés de referência e dispensa armazenar o modelo de referência, reduzindo a memória de *fine-tuning*. A alternativa (a) afirma que o viés é reintroduzido; (c) afirma que ainda são precisas duas cópias; (d) afirma que p_{ref} é mantido.

XII) **(a)** Listando os pares (*in*, *out*) e contando aqueles em que a *loss* do *in* é menor que a do *out*, obtém-se 8 de 16, logo AUC = 0,50. Um atacante que chuta atinge 0,5, então o vazamento está próximo do ideal. A alternativa (b) corresponde ao exemplo do *slide* (ordem diferente); (c) e (d) não correspondem à contagem. Detalhe: com a ordem dada, os *in* ocupam posições $2, 3, 6, 7$ e os *out* posições $1, 4, 5, 8$, e a contagem de pares favoráveis resulta em $8/16$.

Dissertativas.

I) (a) O *oracle* é $\theta^\star = \mathcal{A}(D_r)$, o modelo hipotético retreinado do zero apenas no *retain set*, sem nunca ter visto D_f. Ele é a referência de qualidade porque define o comportamento "correto" de quem nunca leu D_f: o sucesso de θ_u é medido pela proximidade de suas distribuições de saída às de $\theta^\star$ em consultas relacionadas a D_f. Em produção ele não é usado por ser caro de obter.

(b) O fator de aceleração é $184\,000/1 = 184\,000$, ou seja, da ordem de 10^5. O ganho não vem de graça porque o *unlearning* aproximado abre mão da *Forget Quality* exata do *oracle*: o resultado é próximo segundo critérios estatísticos (equivalência distribucional medida por KS, AUC, etc.), não idêntico. Trocamos garantia de esquecimento por custo.

(c) "Aproximado" qualifica o quão perto θ_u fica de $\theta^\star$, e $\theta^\star$ é um modelo que nunca leu D_f, portanto também não responde corretamente perguntas que exigiriam ter lido D_f. Acertar essas perguntas seria, na verdade, evidência de que o conteúdo não foi esquecido. O alvo é a indistinguibilidade distribucional em relação ao *oracle*, não a acurácia sobre D_f.

II) (a) VerbMem (memorização literal) e KnowMem(*forget*) devem ser **baixos**: 0,05 e 0,18 são bons. Utility (KnowMem no *retain*) deve ser **alto**: 0,93 é bom. PrivLeak é uma AUC de *membership inference* cujo alvo é $\approx 0{,}5$: 0,87 está longe do alvo, portanto ruim.

(b) KnowMem(*forget*) baixo mostra que o modelo deixou de responder o que exigia ter lido o **texto** dos livros, enquanto Utility alto mostra que ele preservou o conhecimento do **universo** (FanWiki, personagens, magia). O contraste demonstra que é possível esquecer o texto sem apagar o conceito, que era exatamente o objetivo da separação D_f versus D_r.

(c) AUC = 0,87 significa que, na maioria dos pares (*in*, *out*), o *in-member* (trecho que estava em D_f) ainda tem *loss* sistematicamente menor que um *out-member*, então o modelo continua "reconhecendo" o que deveria ter esquecido. As métricas *deployer-side* mostram que o comportamento visível mudou (não há mais regurgitação literal nem QA correto), mas o vazamento de pertencimento persiste no nível da *loss*: o esquecimento foi superficial, não profundo, e o modelo não é privado.

III) (a) O produto FQ × Utility cai de 0,77 para 0,20 quando $|D_f|$ cresce de 50 para 5000, ou seja, em cerca de duas ordens de grandeza de log o produto cai por um fator $\approx 3{,}9$. A queda é mais que linear no logaritmo de $|D_f|$, logo o método **não é escalável** segundo o critério da aula. Em produção, isso significa que apagar trechos pequenos (capítulos individuais) é fácil, mas apagar conjuntos grandes (livros inteiros) é difícil.

(b) A taxa média de degradação de Utility por pedido é $(0{,}90 - 0{,}41)/4 \approx 0{,}12$ por pedido. Isso indica **sustainability** baixa: a cada pedido o modelo perde utilidade, e a degradação tende a ser sobreaditiva ao longo da sequência. Como o GDPR implica um fluxo contínuo de pedidos, um operador em produção precisa de *sustainability* alta (próxima de manter a Utility estável), o que este método não oferece.

IV) (a) A falha é o *catastrophic collapse*: como $-\mathcal{L}_{\text{NLL}}(D_f, \theta)$ é ilimitada inferiormente, o *ascent* sem freio empurra os *logits* de D_f para $-\infty$ e degrada o modelo em qualquer entrada. A correção mais direta é acoplar o termo de *retain*, usando $\mathcal{L}_{\text{GA}}(\theta) = -\mathcal{L}_{\text{NLL}}(D_f, \theta) + \lambda\, \mathcal{L}_{\text{NLL}}(D_r, \theta)$ com λ típico em $[1, 10]$, que penaliza a perda de utility em D_r. Métodos com perda autorregulada (NPO, SimNPO), em que a sigmoide satura e o gradiente vai a zero quando o item já foi esquecido, evitam o colapso de forma mais robusta.

(b) O IDK é uma **política de resposta**: o modelo aprende a *frase* "*I don't know*" para perguntas, mas o conteúdo memorizado dos livros continua nos pesos. Por isso é vulnerável a *prefix-leak attacks*: quando o atacante força a continuação de um trecho literal em vez de fazer uma pergunta, a memorização vaza. Um método que otimiza diretamente a *log-prob* de D_f (GA com *retain*, NPO, SimNPO) ataca a causa, reduzindo a probabilidade do próprio conteúdo, em vez de apenas trocar a resposta às perguntas.

(c) A propriedade violada é a **robustness** (esquecimento durável), e o ataque é o *relearning attack*: poucos passos de *fine-tuning* numa amostra residual trazem o conhecimento de volta. Entre os métodos vistos, a SimNPO foi reportada como mais resistente a *relearning* que NPO e GA.