Revisão sobre Aumento de Dados

Autor: Matheus Muniz Damasco

Introdução

A Escassez dos Dados Públicos de Texto

No artigo ‘Will we run out of data? Limits of LLM scaling based on human-generated data’, estimou-se que o estoque de textos públicos humanos de alta qualidade atingirá seu ponto médio de esgotamento em 2028, com alta probabilidade de esgotamento completo até 2032 [1].

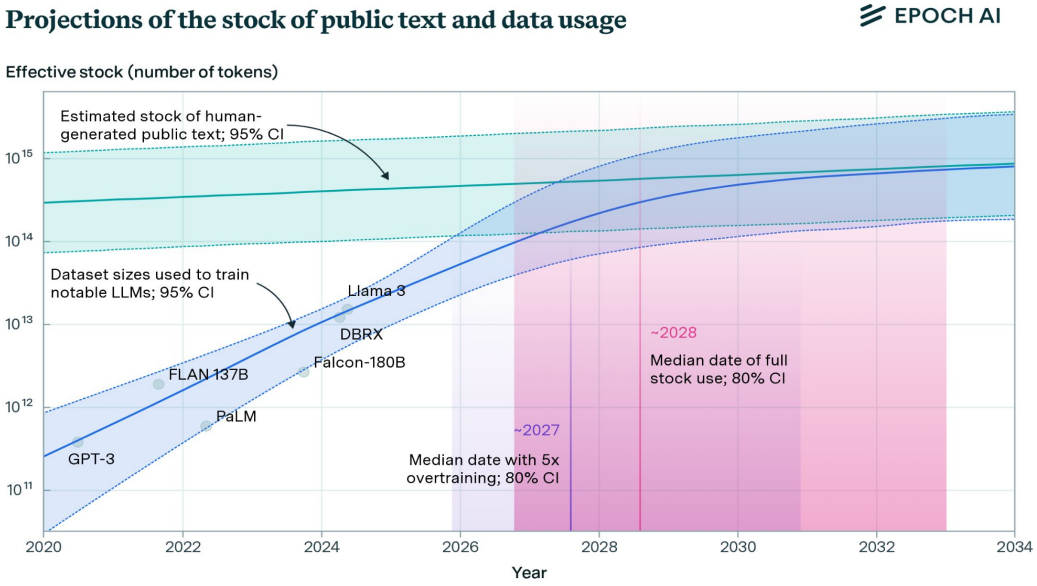


Figura 1: Projeção de estoque efetivo de texto público gerado por humanos e tamanhos de conjuntos de dados usados para treinar LLMs notáveis. Pontos individuais representam tamanhos de conjuntos de dados de modelos notáveis específicos. A projeção de tamanho de conjunto de dados é uma mistura de uma extrapolação de tendências históricas e uma projeção baseada em computação que assume que os modelos são treinados de forma otimizada para computação [1].

A Escassez dos Dados Clínicos e Multilíngues

No artigo recente “Adapting LLMs for the Medical Domain in Portuguese: A Study on Fine-Tuning and Model Evaluation” [2], os autores, ao realizar o fine-tuning de um modelo de LLM, utilizam conjuntos de dados traduzidos para o português brasileiro com o objetivo de desenvolver um assistente médico virtual, especificamente um chatbot especializado em medicina. Idealmente, o modelo seria ajustado com um conjunto de dados nativo, composto por conversas médicas verificadas no contexto brasileiro. No entanto, esse tipo de corpus ainda não está disponível na literatura. Os autores destacam que os conjuntos de dados existentes no contexto brasileiro consistem em registros históricos de interações médicas do século XVI [3] ou em traduções automáticas sem validação profissional [4].

Privacidade com Dados de Treinamento de IA

- **Uso dos dados em plataformas de IA:** Muitas plataformas de IA como ChatGPT, Gemini, Copilot, Grok e outras, coletam dados dos usuários para aprimorar seus modelos. Por padrão essa coleta de dados sempre vem habilitada então desde o primeiro uso caso o usuário não desabilite a opção suas informações podem ser utilizadas para o treinamento dos modelos.
- **Uso dos dados em API de IA:** O uso dos dados via API pelo que busquei parece que funciona da seguinte forma quando a API é paga eles dizem que seus dados são protegidos mas caso você esteja no usando de forma gratuita os dados seram usados.
- **Plataforma menos Transparente:** De todas as plataformas a Meta é a menos transparentes com o uso dos dados do usuário e caso você não esteja na Europa ou Reino Unido suas opções para evitar que seus dados seja usados para treinamento são mínimas.

Direitos Autorais no Treinamento de IA

Várias ações judiciais foram movidas em tribunais da Califórnia e de Nova York desde 2023 contra várias empresas de IA por violação de direitos autorais com base na cópia não autorizada de obras de autores pelas empresas para treinar seus modelos de IA generativa. O mais famoso por enquanto é o processo do Authors Guild Vs Open AI.

A questão central gira em torno de se o uso de obras protegidas para treinamento de IA se enquadra como "uso justo" ou constitui violação de direitos autorais.

Técnicas de Aumento de Dados

Técnicas Básicas de Aumento de Dados

Substituição de sinônimos:

Selecione aleatoriamente "k" palavras em uma sentença que não sejam stop words e substituímos elas pelos seus sinônimos. Para obter sinônimos, podemos utilizar os Synsets do WordNet [5,6].

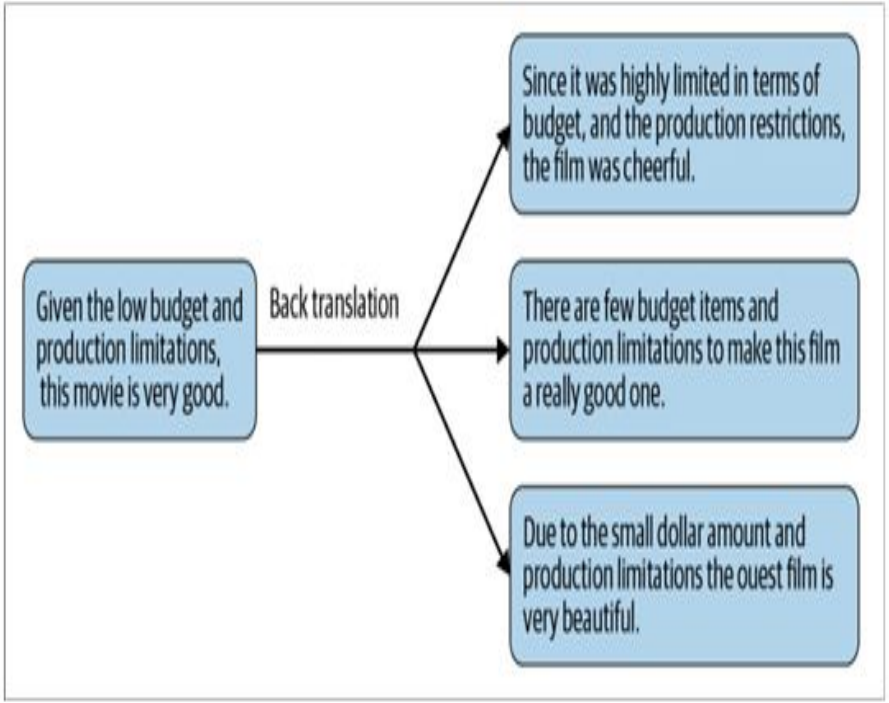
Inversão de Bigram:

Divida a sentença em bigrams. Selecione aleatoriamente um bigrams e inverta sua ordem. Por exemplo: "Eu estou indo ao supermercado." Se escolhermos o bigram "indo ao", ao inverte-lo, a sentença resultante seria: "Eu estou ao indo supermercado."

Técnicas Básicas de Aumento de Dados

Back-Translation:

Considere uma sentença S1 em inglês. Utilizamos uma biblioteca de tradução automática, como o Google Tradutor, para traduzi-la para outro idioma, por exemplo, alemão, resultando na sentença S2. Em seguida, traduzimos S2 de volta para o inglês, obtendo a sentença S3. Observamos que S1 e S3 possuem significados muito semelhantes, mas apresentam variações sutis. Podemos então adicionar S3 ao nosso conjunto de dados. Este método é especialmente eficaz para classificação de texto.



Técnicas Básicas de Aumento de Dados

Substituição de Entidades:

Substitua entidades como nomes de pessoas, locais, organizações, etc… por outras entidades da mesma categoria. Por exemplo, em "Eu moro na Califórnia", substituímos "Califórnia" por "Londres".

Substituição de Palavras Baseada em TF-IDF:

A back-translation pode omitir certas palavras cruciais na sentença. Para lidar com isso, os autores do artigo ‘Unsupervised Data Augmentation for Consistency Training’ em utilizam o TF-IDF [7].

Técnicas Básicas de Aumento de Dados

Adição de Ruído no Dados:

Em muitas aplicações de NLP, os dados recebidos contêm erros ortográficos, principalmente devido às características da plataforma onde são gerados (por exemplo, Twitter). Nesses casos, podemos adicionar um pouco de ruído aos dados para treinar modelos mais robustos. Por exemplo, escolhemos aleatoriamente uma palavra em uma sentença e substituímos por outra palavra com grafia semelhante. Outro tipo de ruído é o problema de "dedo gordo" em teclados móveis. Podemos simular erros de teclado QWERTY substituindo alguns caracteres por seus vizinhos no teclado.

Técnicas Básicas de Aumento de Dados

Snorkel e Snorkel Flow [8]:

O Snorkel foi criado em 2016 na Universidade de Stanford. Trata-se de um framework projetado para automatizar a criação de dados de treinamento rotulados, reduzindo ou eliminando a necessidade de anotação manual. Ele utiliza funções de rotulagem heurísticas, definidas por especialistas, para rotular grandes volumes de dados não anotados de forma programável. Embora não gere dados sintéticos no sentido tradicional, o Snorkel pode ser combinado com técnicas de transformação de dados para enriquecer o conjunto de treinamento. Essa abordagem já foi aplicada com sucesso por empresas como Google, Intel e Stanford Medicine, além de ter resultado em mais de sessenta publicações revisadas por pares, relacionadas a descobertas no uso do Snorkel e inovações em supervisão fraca, aumento de dados, aprendizado multitarefa, entre outros tópicos.

O Snorkel Flow é uma plataforma de machine learning de ponta a ponta, voltada para o desenvolvimento e a implantação de aplicações de inteligência artificial. A plataforma incorpora muitos dos conceitos do projeto original, além de técnicas mais recentes em supervisão fraca, aumento de dados, aprendizado multitarefa, segmentação e estruturação de dados, monitoramento, análise de desempenho, entre outras funcionalidades.

Técnicas Básicas de Aumento de Dados

Easy Data Augmentation (EDA) [9] e NLPAug [10]:

O EDA é uma biblioteca que reúne um conjunto de técnicas simples e eficazes para aumento de dados em tarefas de classificação de texto, sendo especialmente útil em cenários com conjuntos de dados pequenos. As técnicas principais incluem substituição por sinônimos, inserção aleatória, troca de palavras e remoção aleatória — todas aplicadas de forma a preservar o sentido geral das frases originais.

O NLPAug é uma biblioteca mais abrangente voltada para aumento de dados em tarefas de Processamento de Linguagem Natural, tanto em texto quanto em áudio. Ela oferece uma ampla gama de técnicas, como simulação de erros de digitação, substituição de caracteres semelhantes, substituição por sinônimos ou antônimos, inserção, exclusão e troca de palavras, além de métodos mais avançados como back-translation e substituição contextual com modelos pré-treinados como o BERT. Seu objetivo é facilitar a geração de dados sintéticos para melhorar o desempenho e a robustez de modelos de aprendizado de máquina em diferentes domínios.

Técnicas Avançadas de Aumento de Dados

Variational Autoencoders (VAEs) [11]:

Entre os modelos de deep learning, os autoencoders variacionais (VAEs) foram introduzidos em 2013. Eles funcionam aprendendo a compactar os dados de entrada em uma representação menor, conhecida como espaço latente, que é então usada para reconstruir os dados originais. Os VAEs diferem dos autoencoders tradicionais por aprenderem a estrutura subjacente da distribuição dos dados, em vez de apenas memorizar os dados. Isso significa que eles podem criar novos dados semelhantes àqueles nos quais foram treinados, tornando-os ideais para tarefas como geração de imagens, detecção de anomalias e compressão de dados. Os VAEs se tornaram ferramentas essenciais em diversas áreas, incluindo visão computacional e processamento de linguagem natural, devido à sua capacidade de aprender e gerar padrões complexos de dados [12, 13].

Técnicas Avançadas de Aumento de Dados

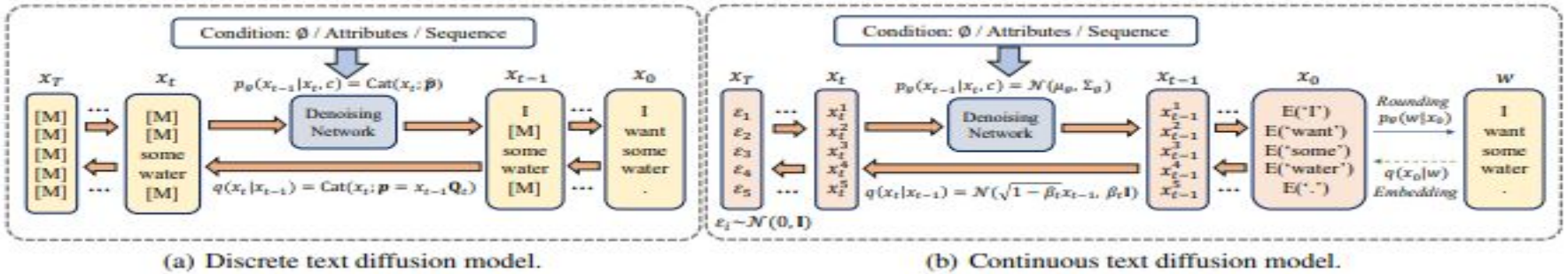
Generative Adversarial Networks (GANs) [14]:

O método introduzido em 2014, conhecido como Redes Adversárias Generativas (GANs), consiste em duas redes neurais que competem entre si: uma Geradora e uma Discriminadora. A rede Geradora tem como objetivo criar dados sintéticos que imitam os dados reais de treinamento, enquanto a Discriminadora busca distinguir entre dados reais e gerados. Durante o treinamento, ambas as redes são aprimoradas continuamente: a Geradora melhora na criação de dados realistas, e a Discriminadora se torna mais eficaz em identificar falsificações. Esse processo competitivo resulta na geração de dados sintéticos cada vez mais convincentes. As GANs têm sido amplamente utilizadas em diversas aplicações, como geração de imagens, aumento de dados e criação de conteúdo original, incluindo arte e música. Essa arquitetura inovadora tem atraído um interesse significativo em pesquisas.

Técnicas Avançadas de Aumento de Dados

Diffusion Models [15]:

Modelos de difusão textual visam recuperar gradualmente um ruído aleatório para um texto desejado, com base nos dados de entrada fornecidos. O ruído inicial pode ser discreto (por exemplo, tokens [MASK]) ou contínuo (por exemplo, ruído gaussiano aleatório), correspondendo ao modelo de difusão discreto ou contínuo. O processo de remoção de ruído depende de uma rede de remoção de ruído parametrizada, geralmente implementada pela arquitetura Transformer [16]. Durante o treinamento, a rede aprende a recuperar os resultados intermediários ruidosos com base nas configurações de programação de ruído, função objetivo e estratégia de condicionamento. Durante a inferência, começando a partir de um ruído aleatório, a rede de remoção de ruído o limpa progressivamente em cada etapa, até produzir o texto-alvo. Observe que, em cada etapa, seguindo o modo de geração não-autoregressivo (NAR), os modelos de difusão textual prevêem todas as variáveis latentes em paralelo.



Técnicas Avançadas de Aumento de Dados

Large Language Models (LLMs) [17]:

Nos últimos anos, os Modelos de Linguagem de Grande Escala (LLMs) emergiram como uma nova abordagem para a geração de conjuntos de dados sintéticos. Modelos como o GPT exemplificam essa capacidade, destacando-se por suas excepcionais habilidades de aprendizado contextual e vasto conhecimento linguístico pré-treinado. O GPT é pré-treinado de forma não supervisionada em grandes volumes de dados da Internet, com o objetivo de prever a próxima palavra em um determinado contexto. Essa estratégia de pré-treinamento permite ao modelo adquirir uma compreensão profunda da estrutura estatística linguística e das relações contextuais, capacitando-o a realizar diversas tarefas de processamento de linguagem natural [18].

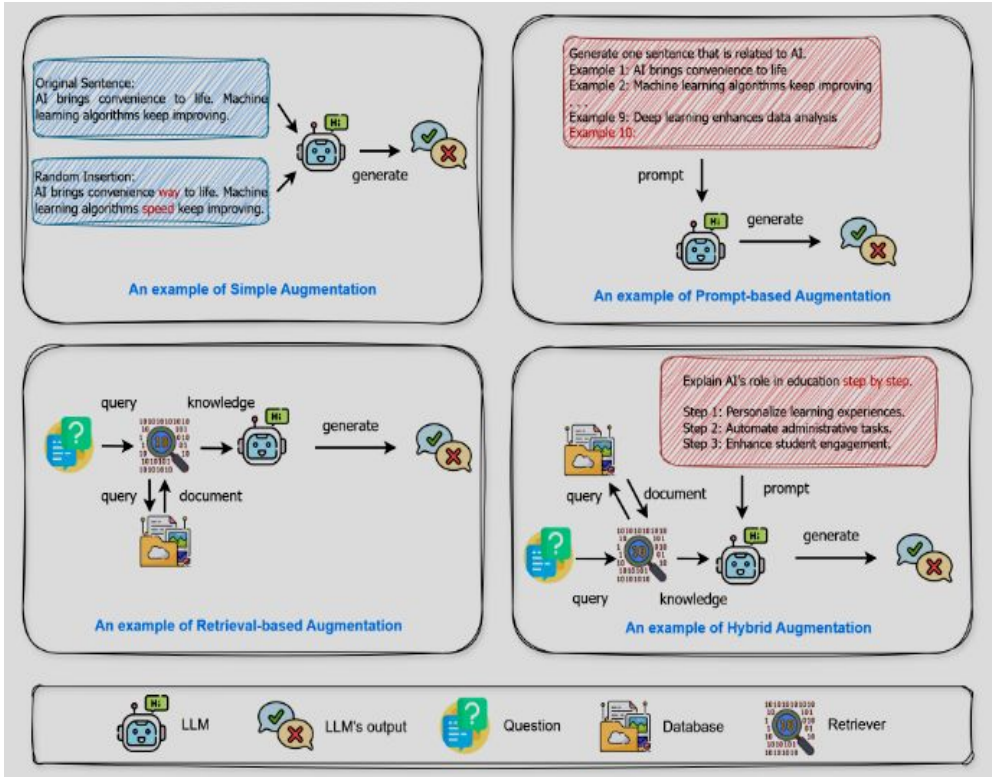
Quatro categorias de técnicas de aumento de dados em LLMs [19]

- Simple Augmentation:

  Aumento Simples é uma técnica inicial de aumento de dados, como substituição de sinônimos, exclusão de palavras e back-translation.

- Prompt-based Augmentation:

  O sucesso do prompt engineering estimulou drasticamente as capacidades dos LLMs [20]. Fornecer aos LLMs prompts cuidadosamente projetados pode fazer com que eles produzam respostas mais humanas. A técnica de Aumento de dados baseado em Prompts melhora efetivamente o desempenho de muitas tarefas, fornecendo prompts relacionados à tarefa ou entre tarefas para orientar os LLMs na geração de dados de alta qualidade.



Quatro categorias de técnicas de aumento de dados em LLMs [19]

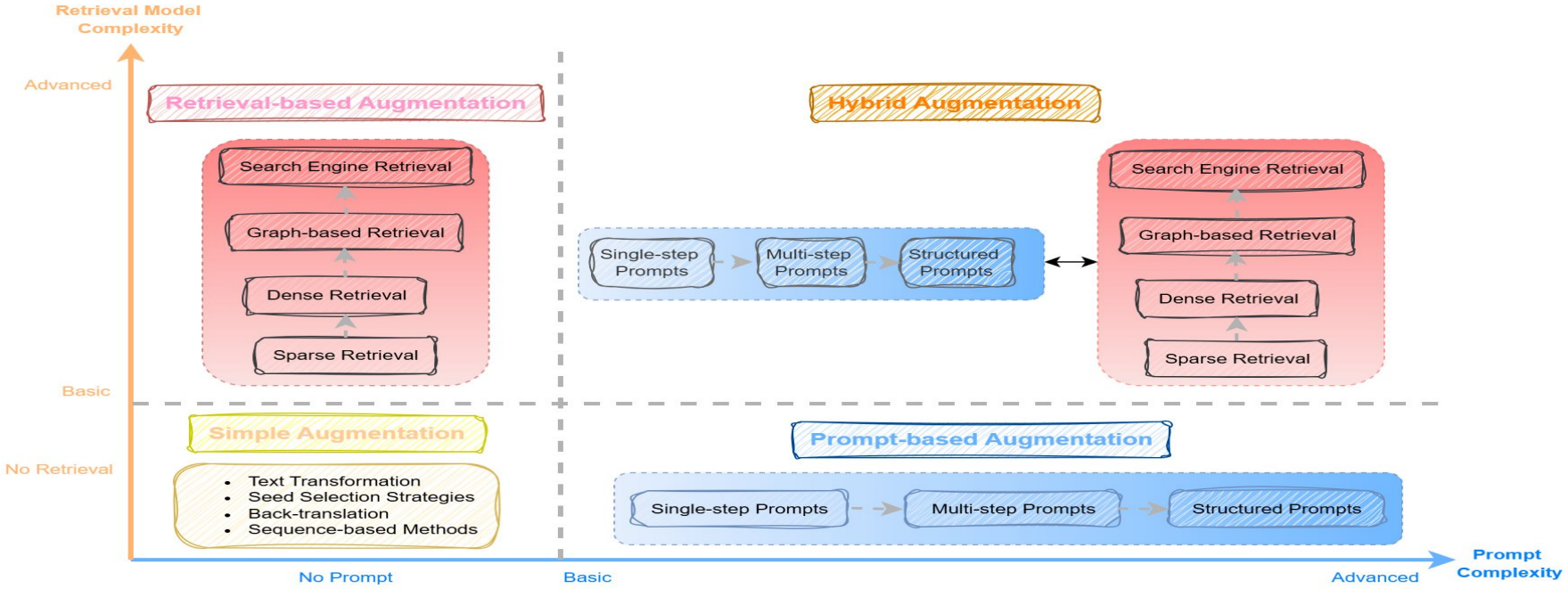
- Retrieval-based Augmentation:

Embora os LLMs tenham demonstrado capacidades satisfatórias em muitas áreas, eles inevitavelmente sofrem com a produção de alucinações e a incapacidade de usar informações externas [28]. O Aumento Baseado em Recuperação supera efetivamente algumas limitações existentes dos LLMs e fornece uma maneira inovadora de aumento de dados, recuperando conhecimento enorme e dinâmico de bases de corpus ou documentos externos [29]. Recentemente, muitos estudos têm utilizado a Geração Aumentada por Recuperação (RAG) [30] para obter informações externas atualizadas em tempo hábil e alcançaram desempenho excepcional em diferentes tarefas. Os métodos de Aumento Baseado em Recuperação podem ser categorizados em Recuperação Esparsa, Recuperação Densa, Recuperação Baseada em Grafos e Recuperação em Mecanismos de Busca.

- Hybrid Augmentation:

A combinação da construção de prompts e componentes de recuperação não apenas estimula os LLMs a produzir conjuntos de dados diversos, mas também explora os recuperadores para obter informações atualizadas em tempo hábil. Pode haver muitas combinações diferentes para construir prompts e realizar a recuperação.

Técnicas de aumento de dados para texto em LLMs



Técnicas de aumento de dados de acordo com a Complexidade do Prompt e de acordo com a Complexidade do Modelo de Recuperação [19].

Comparação das Técnicas Avançadas de Aumento de Dados

Comparações

Modelos	Custo Computacional	Qualidade dos Dados Gerados	Diversidade dos Dados Gerados	Complexidade do Modelo
Variational Autoencoders (VAEs)	Baixo-Médio (2)	Baixo (1)	Baixo-Médio (2)	Baixo-Médio (2)
Generative Adversarial Networks (GANs)	Médio (3)	Baixo-Médio (2)	Baixo (1)	Médio-Alto (4)
Large Language Models (Online - ex: GPT)	Baixo (1)	Alto (5)	Médio-Alto (4)	Baixo (1)
Large Language Models (Local - ex: LLAMA)	Alto (5)	Médio (3)	Médio (3)	Alto (5)
Diffusion Models	Médio-Alto (4)	Médio-Alto (4)	Alto (5)	Médio (3)

Métricas usadas para Avaliação

Tabela 1: Trechos de Estudos sobre Modelos Generativos para Processamento de Linguagem Natural e Métricas Usadas para Avaliação [21].

Estudo	Representatividade	Novidade	Realismo	Diversidade	Coerência
"Generating sentences by editing prototypes" [22]	Perplexity	Qual.	Human eval.	Qual.	Human eval.
"Towards generating long and coherent text with multi-level latent variable models" [23]	NLL, Perplexity	-	Human eval.	Self-BLEU, unique n-grams, 2-gram entropy	Human eval.
"MaskGAN: Better text generation via filling in the_____" [24]	Perplexity	-	Human eval.	Unique 2,3,4-grams	Human eval.

Tabela 2: Trechos de Estudos sobre Modelos para Gerar Dados Sintéticos Médicos e Métricas Usadas para Avaliação [21].

Estudo	Representatividade	Novidade	Realismo	Diversidade	Coerência
"Real-valued (medical) time series generation with recurrent conditional GANs" [25]	MMD, TSTR	NN distance	TRTS	Qual.	-
"Generating multi-label discrete patient records using generative adversarial networks" [26]	Qual.	Disclosure risk	Human eval.	Qual.	Qual.
"Synthesizing electronic health records using improved generative adversarial networks" [27]	K-S test, Dim.-wise stats.	Qual.	ML pred., ARM	Qual.	ARM

Referências

[1] Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., & Hobbhahn, M. (2024). Will we run out of data? Limits of LLM scaling based on human-generated data. arXiv:2211.04325v2 [cs.LG]. Disponível em: https://arxiv.org/html/2211.04325v2

[2] Paiola, P. H., Garcia, G. L., Manesco, J. R. R., Roder, M., Rodrigues, D., & Papa, J. P. (2024). Adapting LLMs for the Medical Domain in Portuguese: A Study on Fine-Tuning and Model Evaluation. arXiv preprint arXiv:2410.00163. Disponível em: https://arxiv.org/pdf/2410.00163

[3] L. Zilio, R. R. Lazzari, and M. J. B. Finatto, “Nlp for historical portuguese: Analysing 18th-century medical texts,” in Proceedings of the 16th International Conference on Computational Processing of Portuguese, 2024, pp. 76–85.

[4] J. R. S. GOMES, “Askdocs: A medical qa dataset,” https://github.com/ ju-resplande/askD, 2020.

[5] Miller, George A. “WordNet: A Lexical Database for English.” Communications of the ACM 38.11 (1995): 39–41.

[6] NTLTK documentation. “WordNet Interface”. Last accessed June 15, 2020.

[7] Xie, Qizhe, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. “Unsupervised Data Augmentation for Consistency Training”. (2019). Disponível em: https://arxiv.org/pdf/1904.12848

[8] Snorkel - Programmatically Build and Manage Training Data. Disponível em: https://www.snorkel.org/ ; https://snorkel.ai/ ;

[9] Wei, Jason W., and Kai Zou. “Eda: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks”, (2019). Disponível em: https://arxiv.org/pdf/1901.11196 ; https://github.com/jasonwei20/eda_nlp

[10] Ma, Edward. nplaug: Data augmentation for NLP. Disponível em: https://github.com/makcedward/nlpaug

Referências

[11] Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. arXiv 2013, arXiv:1312.6114.

[12] Papadopoulos, D., & Karalis, V. D. (2023). Variational Autoencoders for Data Augmentation in Clinical Studies. Applied Sciences, 13(15), 8793. Disponível em: https://doi.org/10.3390/app13158793

[13] Goyal, M., & Mahmoud, Q. H. (2024). A Systematic Review of Synthetic Data Generation Techniques Using Generative AI. Electronics, 13(17), 3509. Disponível em: https://doi.org/10.3390/electronics13173509

[14] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. arXiv 2014. Disponível em: https://dl.acm.org/doi/10.1145/3422622

[15] Li, Yifan; Zhou, Kun; Zhao, Wayne Xin; Wen, Ji-Rong. (2023). Diffusion Models for Non-autoregressive Text Generation: A Survey. arXiv preprint arXiv:2303.06574. Disponível em: https://arxiv.org/abs/2303.06574

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NIPS, pages 5998–6008, 2017. Disponível em: https://arxiv.org/abs/1706.03762

[17] Radford, Alec; Narasimhan, Karthik; Salimans, Tim; Sutskever, Ilya. (2018). Improving Language Understanding by Generative Pre-Training. OpenAI. Disponível em: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

[18] Hao, S., Han, W., Jiang, T., Li, Y., Wu, H., Zhong, C., Zhou, Z., & Tang, H. (2024). Synthetic Data in AI: Challenges, Applications, and Ethical Implications. arXiv preprint arXiv:2401.01629. Disponível em: https://arxiv.org/abs/2401.01629

[19] Chai, Y., Xie, H., & Qin, J. S. (2025) Text Data Augmentation for Large Language Models: A Comprehensive Survey of Methods, Challenges, and Opportunities. arXiv preprint arXiv:2501.18845. Disponível em: https://arxiv.org/abs/2501.18845

[20] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, “Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing,” ACM Comput. Surv., vol. 55, no. 9, pp. 195:1–195:35, 2023.

Referências

[21] Yin, Z., Zhang, Y., Wang, Y., & Wang, W. (2023). A Survey of Synthetic Data Generation for Healthcare. Em 2023 IEEE International Conference on Artificial Intelligence in Medicine (AIM) (pp. 1–8). IEEE. DOI: 10.1109/AIM57408.2023.10122524; Disponível em: https://ieeexplore.ieee.org/abstract/document/10122524

[22] K. Guu, T. B. Hashimoto, Y. Oren and P. Liang, "Generating sentences by editing prototypes", Trans. Assoc. Comput. Linguistics, vol. 6, pp. 437-450, Dec. 2018.

[23] D. Shen, A. Celikyilmaz, Y. Zhang, L. Chen, X. Wang, J. Gao, et al., "Towards generating long and coherent text with multi-level latent variable models", arXiv:1902.00154, 2019.

[24] W. Fedus, I. Goodfellow and M. A. Dai, "MaskGAN: Better text generation via filling in the_____", arXiv:1801.07736, Jan. 2018.

[25] C. Esteban, S. L. Hyland and G. Rätsch, "Real-valued (medical) time series generation with recurrent conditional GANs", arXiv:1706.02633, 2017.

[26] E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart and J. Sun, "Generating multi-label discrete patient records using generative adversarial networks", arXiv:1703.06490, 2017.

[27] M. K. Baowaly, C.-C. Lin, C.-L. Liu and K.-T. Chen, "Synthesizing electronic health records using improved generative adversarial networks", J. Amer. Med. Inform. Assoc., vol. 26, no. 3, pp. 228-241, Mar. 2019.

[28] A. Lazaridou, E. Gribovskaya, W. Stokowiec, and N. Grigorev, “Internet-augmented language models through few-shot prompting for open-domain question answering,” CoRR, vol. abs/2203.05115, 2022.

[29] D. Yang, J. Rao, K. Chen, X. Guo, Y. Zhang, J. Yang, and Y. Zhang, “IM-RAG: multi-round retrieval-augmented generation through learning inner monologues,” in Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024. ACM, 2024, pp. 730–740.

[30] P. S. H. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020.

Referências

[31]

[32]

[33]

[34]

[35]

[36]

[37]

[38]

[39]

[40]