Sumarização clínica com LLMs: avaliação acessível de modelos abertos e proprietários

1st Victor Augusto Hon Fonseca
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
victorahf@ufmg.br

2nd Wagner Meira Jr.
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
meira@dcc.ufmg.br

***Abstract*—In recent years, significant technological advances have been driven by the emergence of LLMs (Large Language Models). Among the sectors that stand to benefit most from these developments, healthcare is particularly noteworthy due to its direct impact on society. More specifically, clinical text summarization has become an increasingly important task, given the large volume of clinical documents generated daily and the value of concise summaries for improving information comprehension and dissemination. In this context, this work proposes a robust and accessible methodology for clinical text summarization while presenting an open weights model with strong potential for addressing this task effectively.**

***Index Terms*—AI, Healthcare, AI in HealthCare, Deep Learning, Open weights model, Natural Language Processing, OpenBioLLM8B, Summarization, Gemini, Gemini-2.5-Flash**

I. INTRODUÇÃO

A. Problema

Recentemente, o campo da tecnologia tem passado por uma notável evolução, sendo as LLMs um dos mais conhecidos exemplos dela.

Nesse sentido, cabe destacar que mais de 80% das informações na área de saúde se encontram em formato não estruturado [1]. Assim, a sumarização de textos clínicos se mostra como uma área que pode influenciar muito o campo da saúde, facilitando a compreensão de informações, por exemplo.

Logo, busca-se aplicar os modelos grandes de linguagem nesse campo, através do modelo OpenBioLLM8B [8], e propôr uma metodologia robusta, mas gratuita para essa aplicação.

B. Motivação

O setor de saúde corresponde a um dos setores com maior capacidade de trazer benefícios, devido ao seu impacto direto na sociedade. Portanto, a sumarização de textos clínicos se mostra muito necessária, considerando que se encontra dentro desse setor, torna o entendimento das informações e seu compartilhamento mais simples e rápido e a maior parte das informações da área de saúde estão em formato não padronizado.

Nesse sentido, é relevante ressaltar também que o modelo "OpenBioLLM8B" se trata de um modelo com grande potencial para a área de resumir textos da saúde, conforme analisado em MSI1 (Monografia de Sistemas de Informação 1), pois, por exemplo, apresenta uma versão maior de 70 bilhões de parâmetros com elevado desempenho nesse e em muitos outros cenários e comparável a LLMs respeitadas pelo mercado e também apresenta por si só notável desempenho em outras áreas de avaliação [8].

A partir disso, demonstra-se a importância de um trabalho voltado para a sumarização de textos clínicos que esteja relacionado ao modelo grande de linguagem OpenBioLLM8B, feito de maneira robusta e acessível.

C. Objetivos

Diante disso, propõe-se a criação de uma metodologia de qualidade e de fácil acesso para avaliação da capacidade de LLMs na sumarização de textos clínicos, extensível a textos de outros domínios. Além disso, busca-se avaliar o desempenho do modelo mencionado nessa metodologia visando a, idealmente, apresentar para a sociedade uma LLM de utilidade para a tarefa.

D. Estrutura do trabalho

Este trabalho está organizado da seguinte forma: a seção 2 corresponde a um Capítulo referencial com identificação de trabalhos correlatos e referencial teórico, a seção 3 corresponde a um Capítulo de contribuição com uma descrição organizada das atividades conduzidas, a seção 4 corresponde a um Capítulo de fechamento com as conclusões e a relação de trabalhos futuros e a seção 5 corresponde às Referências bibliográficas.

II. REFERENCIAL

A. Trabalhos Correlatos

No artigo: "A Metric-Driven Comparative Study of Text Summarization Models: Insights from State-of-the-Art LLMs" [19], publicado na "Springer", realiza-se uma análise de modelos grandes de linguagem que eram, no momento do artigo, considerados estados da arte e já foram muito utilizados, como "ChatGPT-4o", na tarefa de sumarização de textos. Os autores empregam como métricas de análise de desempenho dos modelos "Rouge", "Meteor" e similaridade do cosseno, por exemplo, as quais serviram como inspiração na elaboração da metodologia proposta nesta Monografia.

Já no artigo: "Physician- and Large Language Model-Generated Hospital Discharge Summaries: A Blinded, Comparative Quality and Safety Study" [20], por sua vez, os autores utilizam apenas uma LLM, o "GPT-4 Turbo", e usam, mais uma vez, métricas para medir a qualidade dos resumos gerados, como o "Meteor". Essa fonte serviu como inspiração ao dar maior destaque para o uso de LLMs na sumarização de textos da área da saúde e

também pelas formas automatizadas que usou para analisar o desempenho desses modelos.

Nesse sentido, é relevante mencionar o trabalho "LLM-Driven Online Aggregation for Unstructured Text Analytics" [2]. Embora o artigo analise um tema diferente, ele faz o uso de uma técnica estatística importante, a "Amostragem Estratificada Aleatória Semântica". No atual trabalho foi utilizada uma técnica baseada nela, a "Amostragem Estratificada Aleatória", mas considerando o parâmetro do tamanho do texto, mensurado a partir do número de palavras contidas nele. Como trata-se de uma tarefa de sumarização de textos, o tamanho deles se mostra muito importante, pois está diretamente ligado à aplicabilidade dos modelos na tarefa.

Além disso, cabe destacar que os artigos: "A Study of BFLOAT16 for Deep Learning Training" [12], "Exploring the Impact of Temperature on Large Language Models: Hot or Cold?" [16], "REAL Sampling: Boosting Factuality and Diversity of Open-Ended Generation via Asymptotic Entropy" [17] e "QLORA: Efficient Finetuning of Quantized LLMs" [11] contribuíram na elaboração dos experimentos citados neste trabalho. Os 3 primeiros auxiliaram na escolha de parâmetros de execução dos modelos, no tipo de número que o modelo vai usar para armazenar e processar informações ("bfloat16", parâmetro "torch_dtype"), que temperaturas mais baixas são melhores para tarefas de resumir textos, e que valores do parâmetro "top_p" mais baixos são melhores para essa tarefa. Entretanto, foi escolhido um valor relativamente mais alto (0.6) para não limitar muito o modelo, considerando que um dos modelos LLM usados tem relativamente poucos parâmetros (8 bilhões) [8]. Já o último artigo auxiliou na escolha da forma de quantização que seria utilizada no modelo open weights para reduzir gastos com GPU, com o intuito de deixar o experimento mais acessível, mas ainda preservando muito do desempenho, que foi o QLoRA.

Por fim, este artigo: "Overview of MultiClinSum Task at BioASQ 2025: Evaluation of Clinical Case Summarization Strategies for Multiple Languages: data, evaluation, resources and results" [5], junto com outras informações a respeito presentes nos sites ligados ao dataset [3] [4], contribuíram para a escolha do dataset a ser utilizado, o qual foi, de fato, o "MultiClinSum". Essa escolha se deve à licença mais permissiva do dataset, em comparação com a do dataset anteriormente escolhido ("MIMIC-IV-Ext-BHC"), ao fato de ser adequado para o tema, considerando que relatos de caso clínico (conteúdo do dataset) são muito úteis para o setor de saúde e se aproximam de resumos de alta (conteúdos também muito úteis) e à ligação dele com um centro de referência, o BSC (Barcelona Supercomputing Center).

B. Referencial Teórico

Google Colab corresponde a um ambiente de desenvolvimento que permite a criação e execução de códigos. Ele disponibiliza GPU (Graphics Processing Unit), que são unidades de processamento essenciais para executar mais rapidamente operações feitas por modelos de IA.

Além disso, é importante mencionar o uso da técnica de IC (Intervalo de Confiança) para descobrir um número mínimo adequado de instâncias na avaliação proposta nesta Monografia [22] e que a técnica da Amostragem Estratificada Aleatória com base no número de palavras dos textos corresponde à dividir o dataset em grupos baseando-se nesse parâmetro e escolher de forma aleatória dentro desses grupos as instâncias que serão utilizadas. Isso garante uma avaliação mais representativa e com um parâmetro relevante, ao invés de somente realizar um sorteio aleatório. Também é relevante destacar que o Bootstrap corresponde a uma ferramenta estatística que aumenta a confiabilidade dos resultados gerados em um cenário onde só se tem acesso a uma amostra dos dados. Ele repetidamente pega parcelas dessa amostra e calcula a métrica desejada, retornando ao final um intervalo de onde o verdadeiro valor dela estaria, caso fossem utilizadas também outras amostras dos dados [18].

Nesse sentido, vale mencionar que o parâmetro "do_sample" permite que o modelo escolha outros tokens além do mais provável [13], "device_map" permite que o programa gerencie onde colocar cada parte do modelo na memória [14], "low_cpu_mem_usage" atua para reduzir o uso da memória "CPU" [15], temperatura corresponde a um parâmetro relacionado à criatividade do modelo de IA e "top_p" corresponde a um parâmetro que determina quais tipos de tokens o modelo vai selecionar, podendo ser opções mais prováveis ou se ele pode considerar também outras opções, quanto maior o valor mais opções são consideradas e quanto menor o valor, menos opções são consideradas. Portanto, para uma tarefa de sumarização de textos clínicos onde também será usado um modelo menor de IA, escolheram-se os valores: "True" (com o intuito de não reduzir muito a fluidez gramatical e o vocabulário do modelo), "auto" (para o programa decidir em qual memória alocar cada parte do modelo), "True" (para tornar o experimento mais acessível, reduzindo gastos com memória), 0.3 (para deixar a criatividade mais baixa, mas sem extremos) e 0.6 (para não reduzir muito a fluidez do texto do modelo com valores muito baixos, reduzindo suas possíveis escolhas, e não deixá-lo muito suscetível a alucinações, o que ocorreria com valores muito altos), respectivamente, para os parâmetros mencionados antes.

Por fim, é relevante indicar, de maneira mais detalhada, as métricas que serão utilizadas para avaliação do desempenho do modelo open weights OpenBioLLM8B e do modelo baseline "Gemini-2.5-Flash". Elas são ROUGE-L (avalia qual é a maior sequência de palavras que aparece nos dois textos na mesma ordem) [23], METEOR (mede a semelhança entre textos analisando palavras iguais, palavras com a mesma raiz, sinônimos e ordem das palavras) [24] e Similaridade do Cosseno (transforma as frases em vetores "embeddings" e mede o cosseno do ângulo entre as direções dos vetores, o que gera uma análise da semelhança semântica dos textos) [25]. Também é relevante destacar a estratégia de normalização linear [21], ela foi usada para transformar os valores da similaridade do cosseno (escala entre -1 e 1) para uma escala entre 0 e 1 de maneira proporcional. Isso foi feito com o intuito de padronizar a escala com a das demais métricas e para que o número de amostras necessárias não aumentasse muito (a fórmula do IC usa os valores mínimo e máximo das métricas para calcular o número mínimo de amostras necessário) [22].

III. CONTRIBUIÇÃO

Inicialmente, após as conclusões de MSI1 indicarem que um tema relevante de se analisar seria "Avaliação do OpenBioLLM8B no MIMIC-IV-Note", foi feita uma análise de datasets que fossem ligados ao conjunto mencionado. Então, um forte candidato encontrado foi o "MIMIC-IV-ext-BHC", mas,

após analisar suas condições de uso, suspeitou-se que ele não poderia ser utilizado em ambientes como Google Colab, onde foram realizados os experimentos deste trabalho, o que impediria o uso da GPU gratuita deles e assim, o trabalho proposto.

Assim, encontrou-se o dataset MultiClinSum ligado ao BSC [3] [4] [5], que foi escolhido, pelas razões mencionadas anteriormente, e decidiu-se expandir o tema proposto por MSI1, não somente realizando a avaliação de um modelo com potencial em um dataset relevante, mas propôr uma metodologia robusta e acessível para avaliar LLMs de forma geral na sumarização de textos clínicos.

Posteriormente, com o intuito de oferecer uma percepção mais direta do desempenho do modelo "OpenBioLLM8B", adicionou-se ao experimento o modelo "Gemini-2.5-Flash" como baseline, devido à sua cota gratuita de uso da API, ao fato de já ter sido uma LLM de referência e por ter sido elaborado por uma organização de renome, o Google. Além disso, adicionou-se à metodologia proposta, uma verificação se o número máximo de tokens dos textos a se resumir do dataset excedia o número máximo de tokens da janela de contexto dos modelos para indicar se novas etapas como "chunking" seriam necessárias para o dataset selecionado, o que não foi o caso, estabeleceu-se um limite máximo de geração de tokens dos modelos (1.5 * número máximo de tokens dos textos de resumos de referência), com o intuito de evitar gastos com tokens desnecessários, mas não enviesar muito os modelos e foi feito um tratamento dos dados, retirando instâncias com valores nulos e caracteres desnecessários dos textos e resumos, que poderiam influenciar negativamente na compreensão deles ("\n", por exemplo).

Como próxima etapa, buscou-se por técnicas estatísticas que poderiam auxiliar no processo de amostragem de dados do Dataset. A forma encontrada foi utilizar o Intervalo de Confiança, usando uma confiança de 95% (um dos padrões mais usados) e uma margem de erro de 4% para assegurar um número razoável de amostras, com qualidade, mas sem ser excessivamente grande. Também utilizou-se da amostragem estratificada aleatória com base no tamanho dos textos a se resumir, medindo-se por meio do número de palavras deles, pelas razões mencionadas anteriormente. Após isso realiza-se a ativação dos modelos, com a aplicação da quantização QLoRA e dos parâmetros "do_sample", "device_map", "low_cpu_mem_usage" e "torch_dtype" no OpenBioLLM8B, juntamente com os parâmetros de temperatura e "top_p" mencionados, esses também no modelo gemini-2.5-Flash, todos com os valores mencionados anteriormente. Também são adicionados "prompts" iguais de sistema e de usuário para os modelos.

Como próximo passo da metodologia, foram geradas as respostas dos modelos e calculado o desempenho dos modelos nas métricas Meteor, Rouge-L e Similaridade do Cosseno e o tempo total gasto. Após isso, calcula-se o tempo médio por documento dos modelos, a avaliação de cada uma das métricas para cada resumo, aplicando normalização linear nos valores da similaridade do cosseno para ficarem entre 0 e 1, e a média das notas por métrica. Por fim, calcula-se a faixa de valores possível para a média usando Bootstrap com nível de confiança de 95% e gera-se gráficos.

A metodologia proposta pode ser resumida em:

- **Etapa 1**
  - Pré-processamento dos dados, Checagem e Amostragem
    - Checar número máximo de tokens nos textos para se resumir do Dataset (de acordo com regra da OpenAI [10])
    - Checar número máximo de tokens das janelas de contexto dos modelos [8] [9]
    - Realizar tratamento dos dados
    - Gerar número mínimo de amostras com IC e uso da Amostragem Estratificada Aleatória
- **Etapa 2**
  - Aplicação de quantização e hiperparâmetros
  - Utilização dos modelos na amostra do MultiClinSum
    - Estabelecimento de Gemini 2.5-Flash como modelo de comparação
- **Etapa 3**
  - Medição do desempenho dos modelos por métricas de desempenho (realizando normalização linear das notas da similaridade do cosseno para ficarem entre 0 e 1)
  - Aplicação da técnica de Bootstrap nas notas retornadas pelas métricas

A. Resultados

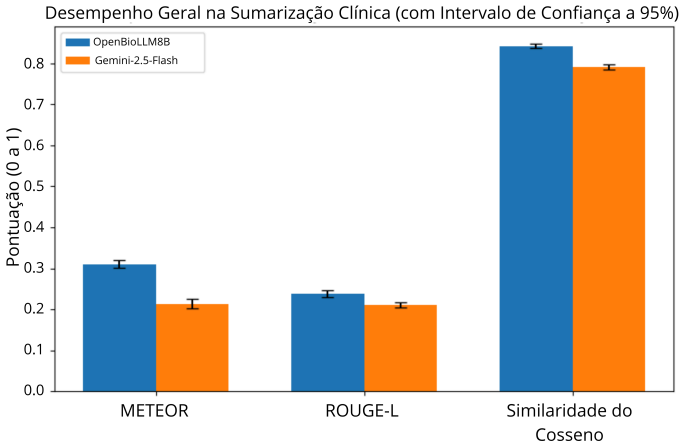

Fig. 1. Gráfico comparativo de avaliações dos modelos pelas métricas METEOR, ROUGE-L e Similaridade do Cosseno

Esse gráfico apresenta as notas médias de cada modelo avaliado em relação às métricas METEOR, ROUGE-L e Similaridade do Cosseno, juntamente com a barra de erro que representa a faixa de valores aos quais elas podem pertencer, fornecido pelo Bootstrap, com confiança de 95%. Pode-se perceber, claramente, que em todas as métricas, mesmo considerando a barra de erro, o modelo OpenBioLLM8B teve um desempenho melhor

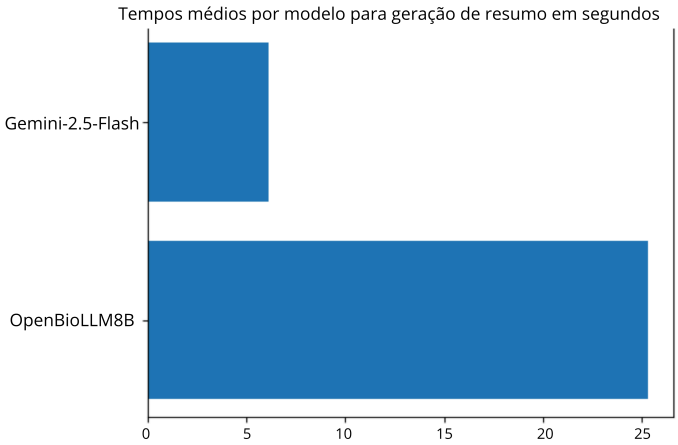

Fig. 2. Gráfico comparativo de tempos gastos pelos modelos

Neste segundo gráfico, mostra-se o tempo médio em segundos de cada modelo para criar a sumarização de um relato de caso clínico (tipo de texto clínico para resumir do "MultiClinSum"). Nota-se que, neste quesito, o modelo Gemini-2.5-Flash teve um desempenho superior.

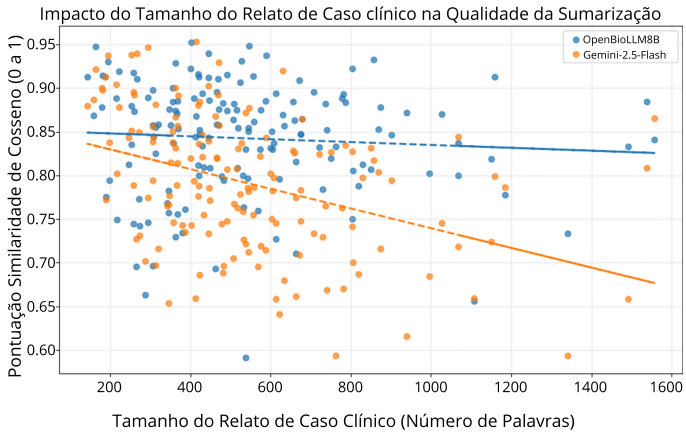

Fig. 3. Gráfico comparativo Tamanho do relato de Caso clínico vs Similaridade do Cosseno para cada modelo

Este gráfico mostra o padrão do valor de similaridade do cosseno (de forma mais direta através de retas de regressão linear) para os resumos gerados por cada modelo, levando-se em consideração também o tamanho do relato de caso clínico em número de palavras. Percebe-se, a partir das retas, principalmente, que, o valor da similaridade do cosseno do modelo OpenBioLLM8B tende a superar o valor correspondente do Gemini-2.5-Flash independentemente do tamanho em número de palavras do relato de caso clínico para a amostra selecionada, o que indica maior qualidade semântica, com seus resumos se aproximando mais semanticamente dos resumos de referência presentes no "MultiClinSum".

IV. CONSIDERAÇÕES FINAIS

A. Conclusões

Diante do exposto, nota-se que uma metodologia robusta e acessível (tanto do ponto de vista de entendimento, quanto de infraestrutura necessária para utilizá-la) é proposta, aplicando estratégias inspiradas em técnicas consolidadas na estatística, como Intervalo de Confiança, Amostragem Estratificada Aleatória e Bootstrap. Além disso, conforme resultados dispostos, evidencia-se a qualidade do modelo OpenBioLLM8B na tarefa de sumarização de textos clínicos, considerando que ele apresentou notas médias maiores que um modelo de LLM que já foi próximo do estado da arte, feito por uma empresa de referência, a Google, em métricas que avaliam mais a estrutura da frase e métricas semânticas, mesmo considerando a barra de erro. Ele também apresentou um padrão de qualidade semântica consistemente maior do que o Gemini-2.5-flash, como pode-se perceber pelo gráfico 3.

Entretanto, percebe-se que o modelo open weights teve um tempo médio por documento maior do que o Gemini-2.5-Flash, isso pode ser explicado pelo fato de que o primeiro é utilizado em uma GPU mais básica, disponível na cota gratuita do Google Colab, enquanto o segundo é utilizado por uma chamada de API.

B. Trabalhos Futuros

Apesar de todas as contribuições propostas pelo trabalho, ele apresenta possíveis pontos de melhoria. Primeiramente, poderia ser interessante repetir o experimento sem limites para o máximo de tokens permitidos para a resposta do modelo.

Por fim, seria interessante repetir o experimento com modelos mais recentes como baseline, como "Claude Opus 4.8", uma das referências em modelos de IA atualmente.

REFERÊNCIAS

[1] M. Sundgren et al., "Unlocking Unstructured Health Data: Scaling eSource-Enabled Clinical Trials," Applied Clinical Trials, vol. 35, no. 1, Feb. 2026. [Online]. Available: https://www.appliedclinicaltrialsonline.com/view/unlocking-unstructured-health-data-scaling-esource-enabled-clinical-trials

[2] C. Hui, W. Lu, Y. Gao, L. Xiong, Y. Wang, and Y. Chen, "LLM-Driven Online Aggregation for Unstructured Text Analytics," arXiv preprint arXiv:2603.08443, Mar. 2026. Available: https://arxiv.org/pdf/2603.08443

[3] "MultiClinSUM: Multilingual Clinical Documents Summarization Shared Task," Barcelona Supercomputing Center, NLP for Biomedical Information Analysis Unit. [Online]. Available: https://temu.bsc.es/multiclinsum/

[4] M. Rodríguez-Ortega, E. Rodríguez-López, S. Lima-López, and M. Krallinger, "MultiClinSum Dataset: Summarization of Clinical Case Reports in English, Spanish, French and Portuguese," Zenodo, version 8, Oct. 13, 2025. doi: 10.5281/zenodo.17341582. Available: https://doi.org/10.5281/zenodo.17341582

[5] M. Rodríguez-Ortega et al., "Overview of MultiClinSum task at BioASQ 2025: evaluation of clinical case summarization strategies for multiple languages: data, evaluation, resources and results," in CLEF 2025, 2025.

[6] Creative Commons, "Attribution 4.0 International (CC BY 4.0) — Legal Code." [Online]. Available: https://creativecommons.org/licenses/by/4.0/legalcode.en

[7] Creative Commons, "Attribution 4.0 International (CC BY 4.0) — Deed." [Online]. Available: https://creativecommons.org/licenses/by/4.0/deed.en

[8] A. Pal (Aaditya Ura) and M. Sankarasubbu, "OpenBioLLMs: Advancing Open-Source Large Language Models for Healthcare and Life Sciences," Hugging Face, 2024. [Online]. Available: https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B

[9] Google Cloud, "Gemini 2.5 Flash — Gemini Enterprise Agent Platform," Google Cloud Documentation. [Online]. Available: https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/gemini/2-5-flash

[10] OpenAI, "What are tokens and how to count them," OpenAI Help Center. [Online]. Available: https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them

[11] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," arXiv preprint arXiv:2305.14314, May 2023. Available: https://arxiv.org/pdf/2305.14314

[12] D. Kalamkar et al., "A Study of BFLOAT16 for Deep Learning Training," arXiv preprint arXiv:1905.12322, May 2019. Available: https://arxiv.org/abs/1905.12322

[13] Hugging Face, "Text generation strategies," Transformers documentation, v4.41.0. [Online]. Available: https://huggingface.co/docs/transformers/v4.41.0/en/generation_strategies

[14] Hugging Face, "Big Model Inference," Accelerate documentation. [Online]. Available: https://huggingface.co/docs/accelerate/usage_guides/big_modeling

[15] Hugging Face, "Models," Transformers documentation, v4.33.0. [Online]. Available: https://huggingface.co/docs/transformers/v4.33.0/main_classes/model

[16] L. Li, L. Sleem, N. Gentile, G. Nichil, and R. State, "Exploring the Impact of Temperature on Large Language Models: Hot or Cold?," arXiv preprint arXiv:2506.07295, Jun. 2025. Available: https://arxiv.org/pdf/2506.07295

[17] H.-S. Chang, N. Peng, M. Bansal, A. Ramakrishna, and T. Chung, "REAL Sampling: Boosting Factuality and Diversity of Open-Ended Generation via Asymptotic Entropy," arXiv preprint arXiv:2406.07735, Jun. 2024. Available: https://arxiv.org/pdf/2406.07735

[18] SciPy Developers, "scipy.stats.bootstrap," SciPy Reference Guide. [Online]. Available: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.bootstrap.htr

[19] S. Gaur, A. Karunik, and S. Naik, "A Metric-Driven Comparative Study of Text Summarization Models: Insights from State-of-the-Art LLMs," in Artificial Intelligence and Speech Technology (AIST 2024), Communications in Computer and Information Science, vol. 2390, A. Sharma and R. Rani, Eds. Cham, Switzerland: Springer, 2025, pp. 192–205. doi: 10.1007/978-3-031-91340-2_16.

[20] C. Y. K. Williams et al., "Physician- and Large Language Model-Generated Hospital Discharge Summaries: A Blinded, Comparative Quality and Safety Study," medRxiv preprint, doi: 10.1101/2024.09.29.24314562, Sep. 2024.

[21] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Waltham, MA, USA: Morgan Kaufmann (Elsevier), 2011.

[22] V. Watts, Introduction to Statistics: An Excel-Based Approach. London, ON, Canada: Fanshawe College Pressbooks, 2022. [Online]. Available: https://ecampusontario.pressbooks.pub/introstats/

[23] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in Proc. Workshop on Text Summarization Branches Out (ACL-04 Workshop), Barcelona, Spain, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013

[24] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," in Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, MI, USA, Jun. 2005, pp. 65–72. [Online]. Available: https://aclanthology.org/W05-0909

[25] H. Steck, C. Ekanadham, and N. Kallus, "Is Cosine-Similarity of Embeddings Really About Similarity?" arXiv preprint arXiv:2403.05440, Mar. 2024. [Online]. Available: https://arxiv.org/abs/2403.05440