Uso de LLMs na Atenção Primária à Saúde

1st Flávio Marcilio de Oliveira
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, MG
flavio.marcilio@dcc.ufmg.br

2nd Marcos André Gonçalves
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, MG
https://orcid.org/0000-0002-2075-3363

***Abstract*—O avanço dos Modelos de Linguagem Ampla (LLM - Large Language Model) permitiu a evolução de aplicações nos mais diversos campos da ciência, inclusive na área da saúde. Muitas pesquisas tem sido realizadas envolvendo diagnósticos de patologias, utilizando os mais variados tipos de dados disponíveis, porém, ainda é rara a utilização de LLMs naquele primeiro contato que os profissionais de saúde tem com os pacientes. Diante do cenário promissor de evolução dos LLMs e da existência dessa lacuna, este trabalho apresenta um estudo da capacidade dos LLMs aplicados na área da atenção primária à saúde, buscando criar uma ferramenta que seja capaz de colaborar com um atendimento individual mais personalizado, com informações realmente úteis para que a comunicação da equipe médica realmente traga valor para os pacientes. Para atingir esse objetivo, uma pesquisa será realizada em todas as etapas de desenvolvimento dos LLMs, buscando compreender os desafios e limitações existentes que impactam na obtenção de um modelo que seja eficiente em cumprir o propósito deste trabalho. A primeira etapa consiste na avaliação da qualidade dos dados disponíveis, buscando compreender os impactos de características como tamanho, legibilidade e gramática do texto nos resultados gerados pelos LLMs. Devido às particularidades da língua portuguesa, será utilizado um pipeline de pré-processamento de dados específico para esse idioma, melhorando a qualidade dos resultados obtidos. Em uma etapa seguinte, é realizado um estudo dos métodos de tokenização que são utilizados para o processamento de linguagem natural, priorizando as duas técnicas principais (técnicas essenciais para o treinamento e uso eficaz de LLMs, como GPT-2 e BERT): BPE (Byte-Pair-Encoding) e WordPiece.**

***Index Terms*—LLM; Atenção primária à saúde; Qualidade de dados; Pipeline de pré-processamento;**

I. Introdução

Os Modelos de Linguagem Ampla (LLM - Large Language Model), principalmente após o surgimento do ChatGPT, deram início a uma nova era de possibilidades e desafios em todos os domínios, inclusive na área da saúde. Várias aplicações nessa área tem sido pesquisadas, principalmente utilizando o ChatGPT. Apesar de não ser treinado em dados médicos, pesquisadores avaliaram a capacidade do ChatGPT em resolver raciocínios de ordem superior no assunto de patologia [15]. Também na área de patologia, um grupo de pesquisadores utlizaram a capacidade do ChatGPT e propuseram um assistente de IA para realizar análise diagnóstica e preditiva em patologia [16]. Os resultados experimentais do assistente mostraram o potencial de aproveitar o modelo de base generativa alimentado por IA para melhorar o diagnóstico de patologias e os processos de tratamento. O desenvolvimento inicial na área de sistemas de apoio à decisão clínica, sistemas inteligentes e avanços baseados em AL/ML na área da saúde iniciaram uma boa plataforma para que o LLM seja integrado e utilizado de forma eficiente no espaço da saúde [8]. O desenvolvimento de assistentes virtuais para auxiliar os pacientes no gerenciamento de sua saúde é outra aplicação importante do ChatGPT na medicina [2].

Diante das promissoras possibilidades de utilização destes modelos, esse trabalho se propõe a investigar a viabilidade de uma nova tecnologia que seja capaz de ampliar as habilidades de Atenção Primária à Saúde de maneira que os profissionais da saúde ofereçam cuidados cada vez mais personalizados para os pacientes, considerando as peculiaridades individuais de cada um.

Assim, para que o objetivo geral seja alcançado, serão perseguidos os seguintes objetivos específicos:

- Investigação, compreensão, estruturação e vetorização de dados históricos relacionados à atendimentos realizados;
- Seleção, pré-treinamento e refinamento de LLMs, priorizando modelos open-source que permitam ter uma segurança maior com dados dos pacientes;
- Criação e avaliação de um fluxo de prompts de interação;
- Integração das soluções em uma ToolChain.

II. Referencial Teórico

A. Large Language Model - LLM

De acordo com [20] "LLMs são enormes modelos de linguagem pré-treinados em grandes quantidades de conjuntos de dados sem ajuste de dados para tarefas específicas" e "modelos ajustados são normalmente modelos de linguagem menores (com até 20 bilhões de parâmetros) que também são pré-treinados e depois ajustados em um conjunto de dados menor e específico da tarefa para otimizar seu desempenho nessa tarefa".

Atualmente, existem diversos LLMs que diferem em suas estratégias de treinamento, arquiteturas e casos de uso. Uma classificação que tem sido adotada na literatura agrupam os modelos em três classes, de acordo com a arquitetura ("Fig. 1"): modelos do tipo codificador-decodificador ("Fig. 1"), modelos somente codificador ("Fig. 1" - Encoder) e modelos somente decodificador ("Fig. 1" - Decoder).

Analisando a evolução dos modelos de linguagem ampla, observa-se que os modelos somente decodificadores

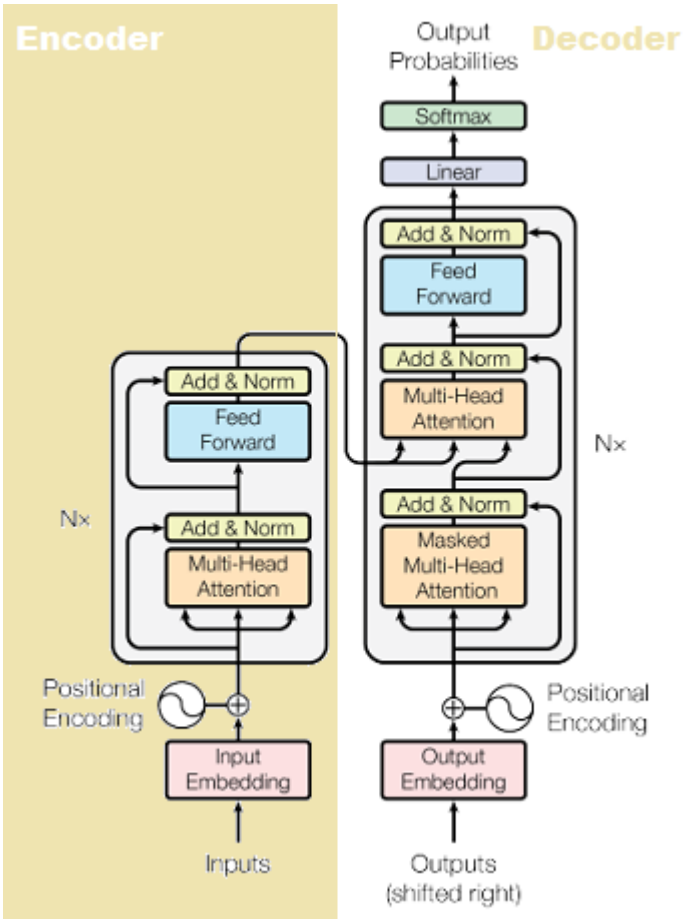

Fig. 1. Arquitetura do LLM (Fonte: Adaptado de [18])

têm gradualmente dominado o desenvolvimento de LLMs. No estágio inicial do desenvolvimento de LLMs, os modelos somente decodificador não eram tão populares quanto os modelos somente codificador e codificador-decodificador. No entanto, após 2021, com a introdução do GPT-3, os modelos somente decodificadores experimentaram um boom significativo. Enquanto isso, após o crescimento explosivo inicial provocado pelo BERT, os modelos somente codificador começaram gradualmente a desaparecer. Os modelos codificadores-decodificadores continuam promissores, pois esse tipo de arquitetura ainda está sendo explorado ativamente e a maioria deles é de código aberto. No entanto, observa-se que os modelos apenas decodificadores apresentam uma maior flexibilidade e versatilidade [20].

Portanto, diante das diferentes arquiteturas disponíveis e, para encontrar o modelo mais adequado para os propósitos deste trabalho, uma etapa a ser desenvolvida será realizar um estudo para compreender as capacidades e limitações envolvidas nos modelos disponíveis atualmente. Para chegar nesse estágio - treinamento dos modelos - será necessário realizar algumas fases anteriores: avaliação da qualidade dos dados disponíveis, pré-processamentos dos dados, modelagem de tópicos.

B. Avaliação da Qualidade dos Dados Disponíveis

A avaliação de qualidade dos dados textuais utilizados neste trabalho foi realizada utilizando como base a metodologia apresentada em [6]. Neste artigo, os autores discutem algumas características que são importantes para avaliar a qualidade textual, que são o tamanho, a estrutura, o estilo e a legibilidade. Neste trabalho utilizou-se as seguintes características:

- **Tamanho:** o tamanho de uma mensagem é avaliado considerando a contagem de frases, palavras e caracteres. De acordo com os autores, mensagens com maior qualidade provavelmente não serão muito curtas nem muito longas.
- **Legibilidade:** a legibilidade é avaliada utilizando estatísticas que quantificam a complexidade de um texto. O método utilizado é o Flesch-Kincaid grade level [10] que gera uma nota geral para avaliação da complexidade do texto. Algumas das outras estatísticas apresentadas aqui são úteis para este cálculo. A fórmula para o cálculo é dada pela equação (1):

$$0.39\frac{\text{total de palavras}}{\text{total de sentencas}} + 11.8\frac{\text{total de silabas}}{\text{total de palavras}} - 15.59 \quad (1)$$

  O retorno é uma nota de acordo com os níveis de ensino americano. Quanto maior, o texto se assemelha à escrita de uma pessoa daquele nível de ensino.
- **Gramática:** Definimos como gramaticalmente correto, a proporção de palavras presentes em um dicionário em português pré-definido e o total de palavras presentes em cada sentença.

C. Pré-processamento dos Dados

A fase de pré-processamento dos dados, normalmente é constituída por algumas etapas básicas, conforme [17]:

- 1. Remoção de stopwords;
- 2. Stemm;
- 3. Remoção de tokens irrelevantes através de técnicas de Seleção de Atributos; e
- 4. vetorização por TF-IDF.

Entretanto, alguns autores tem proposto pipelines estendidos, com inclusão de novas etapas, buscando alcançar melhorias nos estágios de treinamento. Em [1], os autores propõem a inserção de três novas etapas de pré-processamento em um pipeline de classificação de texto melhorando a efetividade e o custo associado com algoritmos de classificação de textos. Contudo, a maioria dos pipelines utilizados e estudados são empregados para a língua inglesa.

Para suprir o gap de pipeline de pre-processamento de texto em português, o trabalho [9] propôs um pipeline com algumas modificações necessárias para a especificidade da língua portuguesa:

- **Etapa 1:** consiste na extração da informação variável de mensagens padronizadas que são utilizadas para estruturar mensagens específicas;
- **Etapa 2:** realiza a normalização de todas as mensagens minúsculo, remoção de acentos e remoção de caracteres que não podem ser representados em ASCII;

- **Etapa 3:** remove as referências aos remetentes e destinatários;
- **Etapa 4:** remove tokens numéricos (e.g. qualquer palavra que contenha números);
- **Etapa 5:** remove URLs com https e links do meet;
- **Etapa 6:** remove o conjunto de pontuação presente na língua portuguesa;
- **Etapa 7:** reduz possíveis repetições de caracteres a no máximo duas ocorrências (e.g. coooocorrência $\Longrightarrow$ coocorrência). Esta etapa também padroniza supostas intensificações de palavras (e.g. boaaa boaaaa $\Longrightarrow$ boaa);
- **Etapa 8:** remove todas as palavras com tamanho menor que três caracteres;
- **Etapa 9:** remove as stopwords, utilizando a biblioteca NLTK (Natural Language Toolkit), com o seu módulo de linguagem em português: "pt_core_news_sm", em conjunto com uma lista adicional de stopwords que inclui expressões de afirmação, saudação, gentileza, etc; Além de uma lista de nomes extraídos da base de dados do IBGE;
- **Etapa 10:** remove o caractere "s" referente ao plural de todas as palavras;
- **Etapa 11:** remove a quebra de linha;
- **Etapa 12:** Lematização dos verbos.

Assim, para os propósitos deste trabalho, este pipeline é o mais adequado e será utilizado na fase de pré-processamentos dos dados.

D. Modelagem de Tópicos

Modelagem de tópicos tenta descobrir relacionamentos entre documentos e tópicos, e entre termos que compõem os documentos e os tópicos, possibilitando a organização e análise de documentos por meio dos tópicos descobertos [7].

O processo de modelagem de tópicos pode ser dividido em três etapas principais: representação de dados, decomposição de tópicos latentes e extração de tópicos. Existem diversos métodos que são empregados para a modelagem de tópicos, porém, para os propósitos deste trabalho, serão avalidos apenas dois: (1) Non-negative Matrix Factorization - NMF [11] e (2) BERTopic [5].

O NMF ("Non-negative Matrix Factorization") utiliza métodos de redução de dimensionalidade para 'matrizes não negativas', semelhante a muitas das abordagens das quais a modelagem de tópicos evoluiu, como redução TF-IDF [12] e LSI/LSA [2].

O BERTopic estende o processo de modelagem de tópico extraindo representação coerente de tópico por meio do desenvolvimento de uma variação baseada em classes do TF-IDF. Mais especificamente, BERTopic gera incorporação de documentos com modelos de linguagem pré-treinados baseados em transformadores, agrupa esses embeddings e, finalmente, gera representações de tópicos com o procedimento TF-IDF baseado em classe [5].

E. Modelos de Tokenização e Vetorização de Dados

O processo de tokenização utilizado em diversos modelos LLM seguem, basicamente, o mesmo pipeline, que inclui as seguintes etapas:

- Normalização
- Pré-tokenização
- Tokenização

Normalização

A normalização é, em poucas palavras, um conjunto de operações de transformação onde uma string bruta torna-se "mais limpa". As operações comuns incluem eliminar espaços em branco, remover caracteres acentuados ou colocar todo o texto em minúsculas. Além disso, a normalização Unicode também é uma operação de normalização muito comum aplicada na maioria dos tokenizadores.

Pré-Tokenização

A pré-tokenização é o processo que consiste em dividir um texto em objetos menores utilizando uma regra de divisão específica, podendo ser espaço em branco, pontuação entre outras. A saída é uma lista de tuplas, com cada tupla contendo uma palavra e seu intervalo na frase original (que é usada para determinar os deslocamentos finais de nossa codificação).

Tokenização

É nessa etapa do pipeline que um modelo de tokenização será treinado em seu corpus para aprender regras de criação de tokens a partir das "palavras" obtidas na etapa de pré-tokenização. A função do modelo é dividir suas "palavras" em tokens, usando as regras que aprendeu. Também é responsável por mapear esses tokens para seus IDs correspondentes no vocabulário do modelo. Existem diferentes técnicas para a tokenização, porém, neste trabalho apresentamos as duas mais comumente utilizadas: BPE e WordPiece.

- BPE (Byte-Pair-Encoding)

Em 1994, Philip Gage propôs um algoritmo geral de compressão de dados, o BPE [4]. Em 2015 o BPE foi adaptado para tokenização [14]. O algoritmo BPE começa com um conjunto de símbolos básicos e combina iterativamente os pares mais frequentes de dois tokens consecutivos no corpus em novos tokens. Para cada fusão, o critério de seleção é baseado na frequência de coocorrência de dois tokens contíguos: o par mais frequente seria selecionado. O processo de mesclagem continua até atingir o tamanho predefinido. Além disso, o BPE em nível de byte (considerando os bytes como os símbolos básicos) tem sido usado para melhorar a qualidade da tokenização para corpus multilíngues (por exemplo, o texto contendo caracteres não ASCII). Os modelos de linguagem representativos com esta abordagem de tokenização incluem GPT-2, BART e LLaMA [21].

- WordPiece

O algoritmo WordPiece foi originalmente proposto pelo Google no desenvolvimento de sistemas de busca por voz [13]. Em 2016, foi usado no *Neural Machine Translation* (NMT), um sistema de tradução automática do Google [19]. Em 2018, foi adotado como algoritmo de tokenização para o BERT [3]. O WordPiece se assemelha ao BPE ao mesclar iterativamente

tokens consecutivos, ao mesmo tempo em que adota um critério de seleção ligeiramente diferente para a mesclagem. Para realizar a fusão, primeiro ele treina um modelo de linguagem e o utiliza para pontuar todos os pares possíveis. Então, a cada mesclagem, ele seleciona o par que leva ao maior aumento na probabilidade de dados de treinamento. Como o Google não lançou a implementação oficial do algoritmo WordPiece, o HuggingFace (https://huggingface.co/) oferece uma medida de seleção mais intuitiva em seu curso online de NLP: um par é pontuado dividindo a contagem de coocorrência pelo produto das contagens de ocorrência de dois tokens no par baseado no corpus de treinamento [21].

III. Metodologia

Na primeira fase desse projeto será feito o levantamento e processamento de todos os dados relacionados com os pacientes, que estão disponíveis, com o objetivo de criar um banco de dados vetorial de forma a viabilizar consultas e adição de novos dados. Um dos desafios visualizados nessa etapa consiste na diversidade de formatos e padrões de dados, além de variações de terminologias e codificações de dados clínicos, os quais deverão passar por um processo de padronização por meio da utilização de estratégias de aprendizagem de máquina. Concomitantemente com esse processo de pré-processamento dos dados, será feito um estudo dos modelos de tokenização e vetorização de dados disponíveis na literatura, avaliando aspectos relacionados à custo e desempenho, com o objetivo de viabilizar a próxima etapa.

Na segunda etapa será realizado um estudo prévio e detalhado de LLMs propostos na literatura, buscando compreender suas arquiteturas, cenários práticos de utilização e, não menos importante, suas limitações. Inicialmente serão priorizados os modelos que sejam open-source pois permitem atender requisitos de privacidade de dados com maior confiabilidade. A meta é selecionar aqueles LLMs que sejam mais aderentes com o escopo do projeto, para, em seguida, realizar o pré-treinamento e o refinamento destes modelos específicos. A clareza nessa definição é crucial, pois impacta diretamente na qualidade das respostas geradas e, consequentemente, na eficácia do sistema. Ainda nessa etapa serão feitas as adaptações necessárias para o estilo de conversação desejado, além de otimizar sua capacidade de produzir respostas contextualizadas.

A terceira etapa do projeto consiste no desenvolvimento de um fluxo eficaz para a criação e avaliação de prompts. A aplicação de técnicas de Engenharia de Prompt representa um desafio, uma vez que existirão incertezas sobre quais estratégias serão mais eficazes na otimização do desempenho do LLM. Para minimizar esses riscos, serão identificados os principais casos de uso do sistema, com a participação essencial de profissionais de saúde.

A quarta e última etapa consiste na implementação de uma ToolChain que integra todas as subsoluções/subprodutos resultantes das etapas anteriores. De forma sucinta, essa ToolChain precisa ser capaz de buscar informações na estrutura de banco de dados criadas para realizar o treinamento e/ou refinamento da(s) LLM(s) selecionada(s), integrando ambas, estrutura de dados e modelos gerados, com prompts gerados e testados na terceira etapa, produzindo o fluxo completo que será utilizado de suporte e apoio na interação entre profissionais de saúde e pacientes.

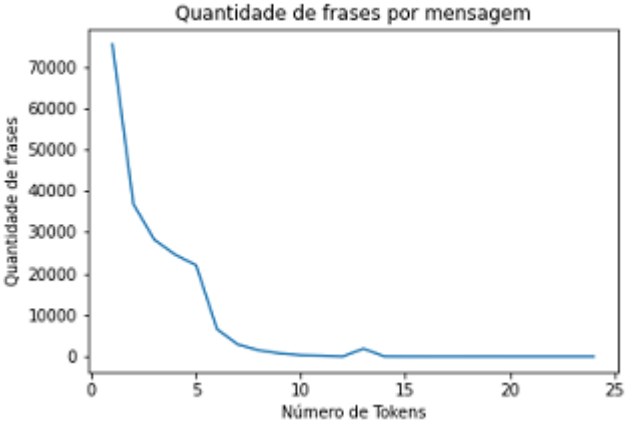

Fig. 2. Quantidade de frases por mensagem.

Fig. 3. Quantidade de palavras por mensagem.

O diferencial da solução proposta está na utilização de contextos extraídos de dados prévios dos sistemas atuais integrados ao LLM por meio de uma ToolChain, resultando em um Modelo de Linguagem Ampla. Dessa forma, será possível gerar prompts contextualizados, com base no histórico do indivíduo, territorialização e determinantes sociais da saúde. A fim de atenuar questões éticas e alucinações, os profissionais de saúde atuarão como curadores do conteúdo produzido pelo LLM (human-in-the-loop).

IV. Desenvolvimento

A. Avaliação da Qualidade dos Dados Disponíveis

Seguindo as métricas apresentadas em [6] foram obtidas as distribuições, para o Tamanho, apresentadas nas “Figs. 2 - 4”:

Outras estatísticas apontadas como importantes pelos autores foram retiradas das mensagens, tais como: tamanho da maior mensagem, média do tamanho das mensagens, número

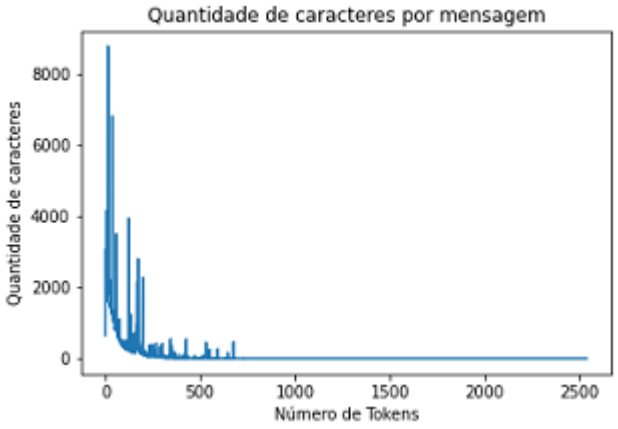

Fig. 4. Quantidade de caracteres por mensagem.

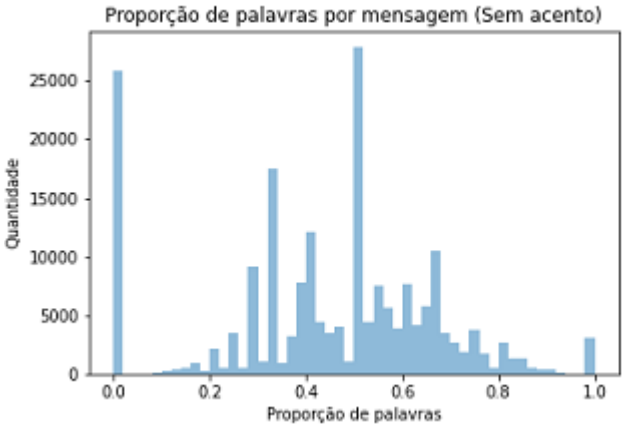

Fig. 6. Proporção de palavras por mensagem (sem acento).

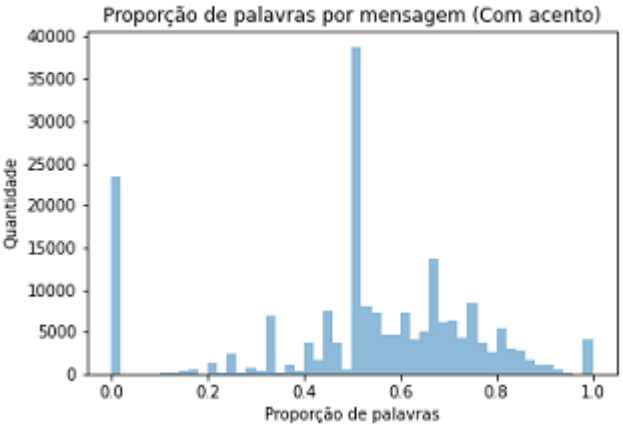

Fig. 5. Proporção de palavras por mensagem (com acento)

de mensagens longas e curtas. Seguindo a metodologia mencionada em [6], definimos como mensagens longas aquelas com 10 palavras a mais que a média, e, por outro lado, as curtas são aquelas com 5 a menos. Segue os valores obtidos por essa análise:

- Maior número de palavras na frase: 121
- Média de tamanho das mensagens: 14.34 palavras
- Número de mensagens longas: 7293
- Número de mensagens curtas: 451734

Ou seja, 98.41% das mensagens possuem em média até 9.34 palavras. As estatísticas de caracteres, frases e sílabas foram calculadas para cada um dos usuário para posteriormente ser utilizadas em análises mais profundas. Algumas delas foram utilizadas aqui para fazer a avaliação de legibilidade.

Com relação à análise gramatical do texto, utilizou-se um dicionário em português, preparado pelo Dicionário "br.ispell", para comparar a proporção de palavras nas mensagens que estão de acordo com o dicionário. A análise foi feita inicialmente considerando acentos ("Fig. 5") e depois os desconsiderando ("Fig. 6").

Considerando ambos resultados obtidos, em conjunto com os anteriores, existe uma forte evidência de que a qualidade dos dados talvez seja um desafio maior do que o previsto para as etapas subsequentes do projeto, especialmente no que tange ao aprendizado e aplicação das LLMs.

B. Pré-processamento dos Dados

Todo pré-processamento realizado sobre as mensagens busca eliminar pequenas variações que podem afetar a contagem de palavras como: "Oi", "oii", "oiii", "oi", "OI", palavras escritas com e sem acento, conjugações verbais como "posso", "poder", "pode", "possa" e muitos outros exemplos. A finalidade desse processo é proporcionar uma análise mais fiel à frequência de palavras relevantes para o domínio do problema,

Antes de realizar qualquer tratamento nas mensagens, são extraídas as informações contidas nas mensagens enviadas automaticamente que são padronizadas, isto é, a parte que não se repete, como nomes ou horários. O objetivo dessa etapa é diminuir a influência causada por essas mensagens que não são relevantes para definição de contexto para a aplicação alvo desse projeto.

Em seguida, é feita a normalização dos dados, retirando acentos e outros caracteres especiais, bem como a transformação de caracteres maiúsculos para minúsculos. Além disso, considerou-se que números não seriam tão interessantes nesta análise, se referindo majoritariamente a datas e horários, além dos ids dos remetentes e destinatários. Também foram retirados espaços e quebras de linha para evitar, pequenos erros de parsing dos dados, isto é, separação das palavras.

Ademais, com o mesmo objetivo de tratar a variância de escrita de certas palavras também foi feito um tratamento para retirada de letras repetidas em uma mesma palavra.

Por fim, a lematização realiza um tratamento que busca encontrar os lemas que compõe as palavras analisadas, ou seja, palavras como "no", "posso" e "possa", são transformadas em "em o", "poder" e "poder", permitindo agrupar ainda mais as palavras semelhantes, utilizadas em contextos semelhantes.

C. Modelagem de Tópicos

A seguir são discutidos e analisados os resultados obtidos usando os métodos NMF e BERTopic.

NMF

Após realizar o pré-processamento citado na seção anterior, obteve-se a representação vetorial das palavras presentes nas mensagens utilizando TF-IDF (*Term Frequency - Inverse Data Frequency*), considerando cada mensagem como um documento.

No uso do modelo NMF, inicialmente foi estabelecido dez tópicos com vinte palavras por tópico e utilizando a função de perda de Kullback-Leibler.

Analisando os tópicos gerados, notou-se a presença de substantivos próprios referentes a nomes pessoais. Assim, para obter uma análise com mais palavras significativas, optou-se por adicionar nas stopwords utilizadas, nomes pessoais populares no Brasil, segundo censo do IBGE. É importante ressaltar que tal tratamento não exclui todos os nomes pessoais, mas apenas os nomes que constarem no arquivo. Além disso, percebeu-se que a maioria dos tópicos é povoada por termos, principalmente verbos e substantivos, relacionados à datas e à saúde, tanto física quanto mental. Notou-se, também, que algumas palavras são frequentes em mais de um tópico, o que, nesse contexto, pode-se justificar pela relevância geral que possuem e, também, pela proximidade dos assuntos retratados nos tópicos.

Diante dessas observações obtidas nessa primeira análise, concluímos que utilizar menos termos em cada tópico é capaz de proporcionar uma análise mais focada e, portanto, mais interessante nesse momento. Portanto, novas análises foram realizadas considerando dez tópicos com apenas dez palavras.

Nessa segunda análise, constatou-se que em alguns tópicos é possível identificar o contexto principal dos documentos, como, por exemplo, o tópico 8 se refere ao reagendamento de um atendimento devido ao cancelamento do mesmo e o tópico 6 é relacionado aos relatos de associados sobre os sintomas ou desconfortos físicos que estão sentindo. Por outro lado, outros tópicos apresentaram termos genéricos demais para determinar com clareza qual o contexto dos documentos daquele tópico. A exemplo disso, é possível citar o tópico 5 que possui apenas verbos.

Para comparar os resultados obtidos com a NMF utilizamos o BERTopic como estratégia de modelagem de tópicos. Assim, verificamos se conseguimos resultados melhores com os dados disponíveis.

BERTopic - 1ª análise

O BERTopic combina técnicas de agrupamento com o modelo BERT (*Bidirectional Encoder Representations from Transformers*) para realizar a identificação dos tópicos. Levando em consideração os resultados obtidos anteriormente, para utilizar tal modelo, definiu-se 10 tópicos, cada um com 10 termos, e configurou-se a linguagem como português.

Uma singularidade do BERTopic é o fato de que, dentre os 10 tópicos identificados, o primeiro, identificado como tópico -1, possuirá apenas termos considerados outliers, isto é, termos que possuem o comportamento atípico de não se encaixarem em nenhum cluster identificado.

Além de modificar o modelo para identificar os tópicos, optou-se, também, por utilizar uma nova distribuição das mensagens considerando um documento como uma sessão, ou seja, um conjunto de mensagens enviadas e recebidas durante um período de tempo específico (definido como um dia). Assim, apesar de diminuir o número de documentos, aumentou-se o tamanho de cada documento.

Considerando as mensagens enviadas, é possível perceber uma melhora na modelagem sendo possível relacionar os assuntos em um maior número de tópicos:

- Tópico 2: orientações em relação à vacinação;
- Tópico 4: cuidado pessoal em relação ao peso;
- Tópico 5: estímulo a atividades relacionadas a descanso e lazer;
- Tópico 7: possivelmente, relacionado ao questionário WHOQOL (*World Health Organization Quality of Life*);

Nos demais tópicos o refinamento ainda não é satisfatório não sendo possível identificar com clareza os assuntos abordados. Destaca-se, nesse sentido, o tópico 6, que possivelmente está relacionado à instruções para videochamadas com colaboradores. Entretanto, termos genéricos ou sem significado aparente ainda surgem.

Em relação às mensagens recebidas, percebe-se que a definição dos tópicos não foi tão consistente quanto nas mensagens enviadas pela empresa. Assuntos claros estão presentes apenas nos tópicos:

- Tópico 0: possivelmente, atraso do associado;
- Tópico 3: comunicação por meios tecnológicos;
- Tópico 4: alimentação do associado;
- Tópico 5: realização do pagamento;

Os demais tópicos possuem termos genéricos. Destaca-se que a qualidade dos textos é um fator que pode influenciar negativamente na modelagem do tópicos, haja vista que uma palavra pode estar grafada de forma errada (ex.: obrahado ao invés de obrigado, no tópico 6) ou até mesmo resumida (ex.: pfvr ao invés de por favor, como aparece no tópico 7). Isso dificulta o pré-processamento do texto e, na modelagem de tópicos, permite que tais termos apareçam como se fossem relevantes, mesmo que não agreguem sentido ao tópico em questão. Importante ressaltar que a qualidade das mensagens terá um impacto crucial na execução das estratégias baseadas em LLM.

Focando a análise na similaridade dos tópicos, mapeou-se a distância entre tópicos no espaço bidimensional para enxergar como esses tópicos se relacionam (“Fig. 7”).

Na distância entre tópicos, cada círculo representa um tópico, tornando possível visualizar no ambiente 2D seus tamanhos e os termos em cada tópico. Em ambos os mapas ocorre a sobreposição de dois ou mais tópicos em um mesmo espaço, o que explica palavras repetidas em mais de um tópico e evidencia, ainda mais, a proximidade entre alguns tópicos. Para quantificar tal proximidade utiliza-se a matriz de similaridade (“Fig. 8”).

Fig. 7. Distância entre tópicos - Mapeamento 2D

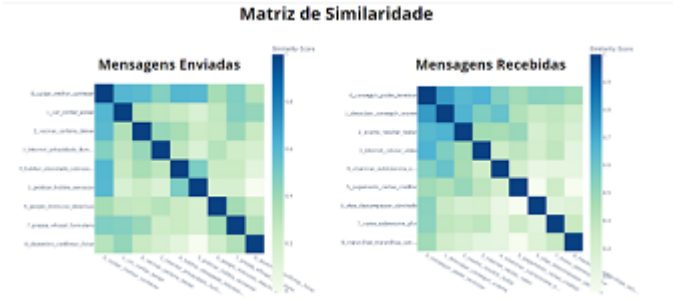

Fig. 8. Matriz de similaridade dos tópicos

Nessa matriz, o grau de similaridade é quantificado pela cor. Nesse sentido, é possível perceber que alguns tópicos genéricos possuem uma alta similaridade, como é o caso dos tópicos 0 e 1 nas mensagens recebidas. Todavia, tópicos bem definidos também podem apresentar uma relação próxima, como é o caso dos tópicos 4 e 5 das mensagens enviadas. Enfim, nesta primeira análise, o BERTopic consegue distinguir melhor os tópicos presentes nas mensagens do que o NMF e possui vantagens claras em relação às análises dos resultados obtidos, porém ainda possui tópicos genéricos não tão definidos no resultado final. Vale lembrar que tal modelo, em geral, é projetado para o uso com textos em inglês e ao utilizá-lo com textos em português os resultados podem sofrer com as diferenças entre essas linguagens. Outro fator importante é que, a partir dos resultados obtidos, existe uma forte evidência de que a qualidade dos dados talvez seja um desafio maior do que o previsto nas etapas subsequentes do projeto, especialmente no que tange ao aprendizado e aplicação das LLMs.

BERTopic - 2ª análise

A modelagem de tópicos utilizando o BERTopic possui natureza hierárquica, isto é, tendo definido o número de tópicos, é possível aproximar a hierarquia entre tópicos utilizando a matriz de tópicos-termos (c-TF-IDF). Nesse sentido, a similaridade entre tópicos é quantificada pela distância entre suas representações c-TF-IDF.

Nessa segunda análise, melhorou-se o conjunto de stopwords (definidas no NLTK), adicionando expressões de afirmação, saudação e agradecimento antes não identificadas, e aumentou-se a quantidade de tópicos para vinte a fim de

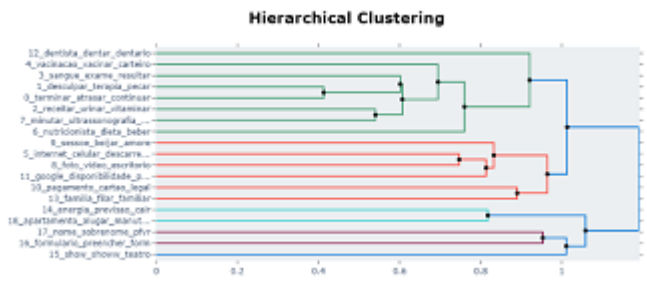

Fig. 9. Dendograma mostrando o agrupamento hierárquico entre os tópicos.

visualizar melhor a relação hierárquica entre tópicos.

Nas mensagens enviadas, a modelagem não apresentou melhorias significativas devido a presença de mensagens padronizadas. A presença dessas mensagens gera clusters com termos genéricos e confusos sem uma definição de tópico. Apesar disso, ainda é possível obter tópicos concisos, como por exemplo, o tópico 18, relacionado à saúde bucal e o tópico 10, relacionado a feriados.

Já na modelagem das mensagens recebidas, testou-se a personalização de alguns parâmetros com o objetivo de obter resultados melhores. A seguir, explica-se quais mudanças foram feitas e quais suas influências nos resultados.

- **reduce_frequent_words:** esse parâmetro é definido, por padrão, como *false* no modelo de c-TF-IDF. Ao configurá-lo como *true*, o modelo reduz a importância de palavras frequentes em todos os tópicos utilizando a raiz quadrada da frequência do termo;
- **bm25_weighting:** também, por padrão, *false* no modelo de c-TF-IDF. Quando configurado como *true*, muda a medida de ponderação. Segundo a documentação do BERTopic, em conjuntos menores, essa mudança na medida de ponderação contribui para a redução de stopwords ainda presentes.

Percebe-se que, apesar de existirem tópicos ainda pouco significativos, novos tópicos interpretáveis surgiram. No gráfico 'Hierarchical Clustering' ("Fig. 9"), os pontos pretos sinalizam a representação de tópico possível naquele nível de hierarquia. Em suma, essa representação facilita a visualização da proximidade de alguns tópicos. Por exemplo, os Tópicos 5 e 8, ambos possuem termos relacionados a comunicação via aparelhos tecnológicos e, empiricamente, percebe-se a conexão entre eles, a qual é comprovada hierarquicamente. Também, os Tópicos 16 e 17, os quais, separadamente, não produzem tanto significado, porém, em conjunto, pode-se inferir a ligação com o preenchimento de questionários.

É perceptível que a quantidade de tópicos influencia na qualidade dos tópicos gerados. Ao mesmo tempo que novos tópicos interessantes são visualizados, tópicos com termos embaralhados surgem.

BERTopic - 3ª análise

Nessa análise, avaliou-se os tamanhos dos clusters gerados até então. Por padrão, o algoritmo de clusterização utilizado pelo BERTopic define que cada grupo deve ter, no mínimo,

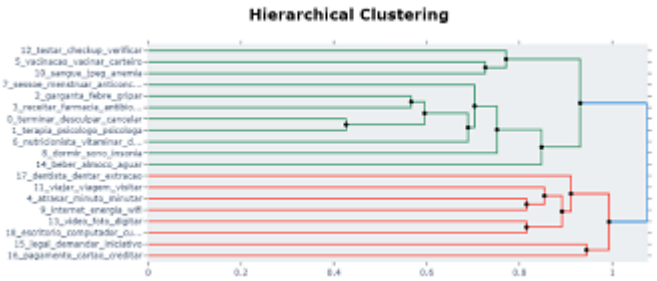

Fig. 10. Agrupamento hierárquico entre os tópicos - Mensagens recebidas.

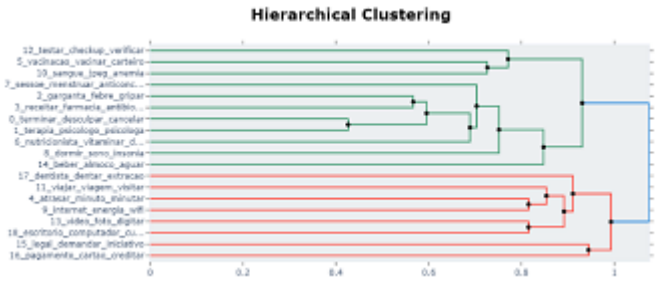

Fig. 11. Agrupamento hierárquico entre os tópicos - Mensagens enviadas.

dez documentos para ser considerado um cluster. Entretanto, percebeu-se, nos resultados obtidos anteriormente, que os últimos tópicos gerados em uma modelagem possuem uma qualidade menor. Assim, para mitigar esse efeito, aumentou-se para trinta o número mínimo de documentos necessários em um grupo para que ele fosse considerado um cluster e, portanto, mais sessões são necessárias para que um grupo fosse realmente considerado um tópico.

A princípio, nas mensagens recebidas, a melhora nos tópicos gerados foi significativa. Agora, os tópicos são mais claros e concisos, possibilitando um estudo mais eficaz dos assuntos tratados nas sessões. Percebe-se, por exemplo, que questões relacionadas à saúde física dos associados são evidenciadas na maioria dos tópicos, porém com focos diferentes, como vacinação, alimentação, sono, sintomas relatados, medicação e exames. Além disso, a saúde mental e psicológica dos associados claramente está associada ao Tópico 1 - o segundo maior tópico nessa modelagem - o que evidencia o quanto essa área é citada nas sessões.

Por outro lado, questões rotineiras também possuem destaque nos tópicos, como o atraso ou o cancelamento de uma consulta, as formas de manter contato digitalmente e as dificuldades enfrentadas nesses contatos, o pagamento de mensalidades e, até mesmo, os lazeres, trabalhos e opiniões dos associados.

O gráfico apresentado na “Fig. 10” permite perceber uma relação próxima entre alguns tópicos. A exemplo disso, é possível citar os tópicos 0 e 1 que, juntos, permitem inferir que os cancelamentos estão diretamente relacionados com as sessões de psicoterapia. Além disso, nos tópicos 4 e 9, relaciona-se os atrasos com instabilidades na conexão. Enfim, a construção hierárquica gerada proporciona uma relação entre tópicos baseada na similaridade e, dessa forma, a interpretação dos tópicos gerados se torna mais completa.

Já nas mensagens enviadas, após o processo de extração das mensagens padronizadas, identificou-se novos tópicos. Nessa modelagem, também aumentou-se o tamanho mínimo necessário de documentos para compor um cluster.

Destaca-se que os tópicos encontrados nas mensagens enviadas condizem com os tópicos encontrados nas mensagens recebidas, o que é coerente, dado que uma determinada sessão entre um colaborador e um associado possui um determinado tópico. Nesse sentido, tópicos relacionados à saúde física e mental dos associados possuem destaque, assim como questões rotineiras relacionadas à comunicação. Além disso, evidencia-se que os lazeres dos associados também são um tópico frequente na troca de mensagens. Por fim, a “Fig. 11” retrata a hierarquia entre os tópicos citados.

Analisando o gráfico apresentado na “Fig. 11”, a proximidade entre alguns tópicos fica mais evidente. A exemplo disso, os tópicos 14 e 16, ambos são relacionados à videochamadas, seja em relação à forma de sua realização ou a problemas que podem surgir. Também, os tópicos 15 e 17, claramente relacionados às atividades de lazer, estão próximos hierarquicamente. Como já citado, entender a hierarquia dos tópicos melhora a interpretação dos mesmo e, por conseguinte, permite uma estruturação mais completa dos assuntos abordados nas sessões.

Em conclusão, no processo de caracterização das mensagens entre colaboradores e associados, a modelagem de tópicos foi uma ferramenta essencial para a separação entre mensagens que realmente abordam os temas das conversas e as mensagens que exprimem pouco significado, as quais não deveriam possuir destaque. Assim, foi possível evidenciar quais tópicos são relevantes para o desenvolvimento de atividades futuras. Mais do que isso, observamos que a qualidade dos textos dos associados não foi um problema para a geração dos tópicos, um indício que talvez também não seja para as próximas etapas do projeto.

V. Conclusões

Para esta primeira fase do projeto foram realizadas tarefas importantes para entendermos melhor os dados com os quais estamos trabalhando e também a parte do negócio que está diretamente relacionada ao projeto. Todo o levantamento e as descobertas obtidas aqui serão utilizadas para direcionar decisões nas próximas fases.

Durante o desenvolvimento dessa etapa do projeto foi possível identificar pontos que poderiam impactar significativamente a execução das tarefas posteriores, demonstrando a importância dessa etapa de caracterização. Por exemplo, observamos a ausência de identificadores únicos das sessões, o que dificultava delimitar, dentro de um contexto, as mensagens trocadas, e que poderia ter impacto significativo no uso da LLM nas fases futuras. Nesse caso, definimos e implementamos uma estratégia para delimitar essas sessões. De forma

simplificada, uma sessão é terminada quando não há interação entre associado e colaborador por mais de 24 horas. Outro exemplo de questão que pode ser identificada e resolvida nessa etapa do projeto foi a lista de mensagens templates ou padronizadas, que foi sendo conhecida e aprimorada nesta fase.

Com relação às próximas etapas do projeto, é importante salientar que já está em pleno andamento e adiantada. As estratégias de padronização precisaram ser implementadas para que pudéssemos realizar a parte de modelagem de tópicos, bem como o processo de padronização de sessões, conforme previamente discutido nesse relatório. Nesse ponto, o que resta seria aprimorar o processo usado de padronização e, eventualmente, incluir novos passos caso as etapas posteriores apontem para essa direção. Com relação às API, o processo de entrada já está pronto, pois foi necessário para pudéssemos ter acesso aos dados usados na elaboração da caracterização. Já iniciamos o processo de definição das saídas e documentação de todo o processo. Por fim, com relação à terceira macro-entrega, também já iniciamos o processo de estudo de LLMs abertas disponibilizadas na literatura.

REFERENCES

[1] Washington Cunha, Sérgio Canuto, Felipe Viegas, Thiago Salles, Christian Gomes, Vitor Mangaravite, Elaine Resende, Thierson Rosa, Marcos André Gonçalves, and Leonardo Rocha. Extended pre-processing pipeline for text classification: On the role of meta-feature representations, sparsification and selective sampling. *Information Processing & Management*, 57(4):102263, 2020.

[2] Tirth Dave, Sai Anirudh Athaluri, and Satyam Singh. Chatgpt in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Frontiers in artificial intelligence*, 6:1169595, 2023.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] Philip Gage. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38, 1994.

[5] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[6] Daniel Hasan Dalip, Marcos André Gonçalves, Marco Cristo, and Pável Calado. Automatic quality assessment of content created collaboratively by web communities: a case study of wikipedia. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, pages 295–304, 2009.

[7] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, 1999.

[8] Mohd Javaid, Abid Haleem, and Ravi Pratap Singh. Chatgpt for healthcare services: An emerging stage for an innovative perspective. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 3(1):100105, 2023.

[9] Antônio Pereira De Souza Júnior, Pablo Cecilio, Felipe Viegas, Washington Cunha, Elisa Tuler De Albergaria, and Leonardo Chaves Dutra Da Rocha. Evaluating topic modeling pre-processing pipelines for portuguese texts. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, pages 191–201, 2022.

[10] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Institute for Simulation and Training. 56.*, 1975.

[11] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *nature*, 401(6755):788–791, 1999.

[12] Gerard Salton. Some research problems in automatic information retrieval. In *Acm sigir forum*, volume 17, pages 252–263. ACM New York, NY, USA, 1983.

[13] Mike Schuster and Kaisuke Nakajima. Japanese and korean voice search. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5149–5152. IEEE, 2012.

[14] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

[15] Ranwir K Sinha, Asitava Deb Roy, Nikhil Kumar, and Himel Mondal. Applicability of chatgpt in assisting to solve higher order problems in pathology. *Cureus*, 15(2), 2023.

[16] Yuxuan Sun, Chenglu Zhu, Sunyi Zheng, Kai Zhang, Zhongyi Shui, Xiaoxuan Yu, Yizhi Zhao, Honglin Li, Yunlong Zhang, Ruojia Zhao, et al. Pathasst: Redefining pathology through generative foundation ai assistant for pathology. *arXiv preprint arXiv:2305.15072*, 2, 2023.

[17] Alper Kursat Uysal and Serkan Gunal. The impact of preprocessing on text classification. *Information processing & management*, 50(1):104–112, 2014.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[19] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[20] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32, 2024.

[21] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.