Análise do Modelo TrafficGPT no Contexto de Ataques Zero-Day

Aluna: Ana Luiza Macêdo dos Santos
Departamento de Ciências da Computação (DCC)
Universidade Federal de Minas Gerais (UFMG)
analuizamacedo@dcc.ufmg.br

Orientadora: Profª. Dra. Michele Nogueira Lima
Departamento de Ciências da Computação (DCC)
Universidade Federal de Minas Gerais (UFMG)
michele@dcc.ufmg.br

***Resumo*—Este estudo investiga a confiabilidade do modelo *TrafficGPT*, uma abordagem baseada em *Large Language Model(LLM)*, para classificação de tráfego criptografado, migrando do cenário tradicional de closed set (em que todas as classes são conhecidas no treinamento) para o de open set (em que o modelo pode encontrar classes nunca vistas e deve ser capaz de identificá-las como desconhecidas). Com foco na detecção de ataques *zero-day*, isto é, ataques que exploram vulnerabilidades até então desconhecidas e que frequentemente se valem de tráfego criptografado, ocultando sua carga maliciosa enquanto mantêm a comunicação aparentemente legítima. Este trabalho analisa a relação entre a confiabilidade do GPT-2 (base do *TrafficGPT*) e a classificação open set, visando contribuir para a redução da efetividade desses ataques. Embora o *TrafficGPT* apresente desempenho promissor na identificação de padrões de tráfego conhecidos e desconhecidos, sua arquitetura exibe superconfiança estrutural, frequentemente gerando previsões excessivamente confiantes mesmo para dados fora da distribuição de treinamento.Para mitigar essa limitação, aplica-se a técnica de *Temperature Scaling*, que recalibra as probabilidades de saída ajustando a escala dos *logits* por meio do parâmetro de temperatura T, otimizado via algoritmo L-BFGS-B. O modelo foi treinado exclusivamente com tráfego benigno do conjunto CIC-IDS2017 e avaliado com a porção contendo ataques simulados, emulando assim um ambiente sob ataque *zero-day*. Os resultados indicam que, embora o *Temperature Scaling* tenha evidenciado a necessidade de calibrar as confiança do modelo, ele não impactou significativamente o desempenho do classificador K-LND em cenário de *open set* — o qual opera com base em medidas de distância geométrica, e não em saídas probabilísticas. Isso sugere que melhorias na detecção em *open set* podem exigir ajustes não apenas na calibração de confiança, mas também nas representações vetoriais internas do modelo e nas métricas de distância empregadas.Por fim, o trabalho ressalta a importância da confiabilidade e da calibração de modelos baseados em LLMs para aplicações críticas de cibersegurança, visando uma detecção mais robusta de ataques *zero-day*.**

***Index Terms*—*TrafficGPT*, LLM, Open Set, ataques zero-day, calibração de confiança, Temperature Scaling, L-BFGS-B, K-LND, Logits**

I. Introdução

A evolução constante das ameaças cibernéticas tem imposto desafios crescentes aos mecanismos de defesa de redes. Entre essas ameaças, os ataques *zero-day* exploram vulnerabilidades ainda não reveladas ao público [2]; enquanto essa falha permanece oculta, não é possível corrigir o software vulnerável, nem detectar a exploração com mecanismos tradicionais de assinatura, tampouco por modelos de aprendizado de máquina supervisionados que assumem conhecimento completo das classes, tornando a defesa extremamente difícil. Em ambientes complexos e dinâmicos, a identificação de tráfego malicioso torna-se ainda mais crítica quando o tráfego está criptografado, reduzindo drasticamente a eficácia de técnicas que dependem de conhecer previamente o conteúdo dos pacotes que trafegam nas redes.

Como base para esta investigação, adotou-se o artigo *"TrafficGPT: An LLM Approach for Open-Set Encrypted Traffic Classification"* [1], que utiliza o modelo de linguagem GPT-2, um *Large Language Model* (LLM), para classificar tráfego criptografado mesmo em cenários em que os componentes dos pacotes de dados sejam desconhecidos pelo *TrafficGPT*, ou seja, que demandem uma classificação open set. O modelo é capaz de reconhecer e classificar semanticamente o tráfego como inofensivo (também compreendido como "legítimo" ou "conhecido") ou malicioso (entendido como "inédito", "fora do conjunto" ou "potencialmente suspeito").

Entende-se que a capacidade do modelo ser uma ferramenta de prevenção de ataques *zero-day* tem relação direta com a necessidade de operar além do paradigma closed set, no qual os classificadores assumem que todas as classes possíveis estão presentes no treinamento. Como destacado pelos autores [1], abordagens tradicionais tentam contornar essa limitação treinando com classes "desconhecidas porém benignas" - as chamadas *"background class*, todavia esses métodos apresentam limitações próprias, pois exigem amostras benignas(desconhecidas porém benignas) durante o treinamento, contudo não há garantias de que tais amostras representem todos os exemplos que o classificador do modelo poderá encontrar no futuro.

Consequentemente, o cenário de cibersegurança demanda classificadores capazes de lidar explicitamente com o regime *open set*, identificando corretamente amostras pertencentes às classes conhecidas *(closed set)* enquanto rejeitam, de forma confiável, qualquer tráfego oriundo de classes não vistas durante o treinamento, o que inclui consequentemente ataques *zero-day*. Diante desse cenário, os modelos que contenham a classificação *open set*, têm sido explorados como alternativa promissora.

Haja vista o que foi desenvolvido na Monografia em Sistemas de Informação I (MSI I), este trabalho propõe-se a aprofundar a análise do paradigma de open set recognition, no qual o modelo deve distinguir tráfego legítimo conhecido de instâncias inéditas

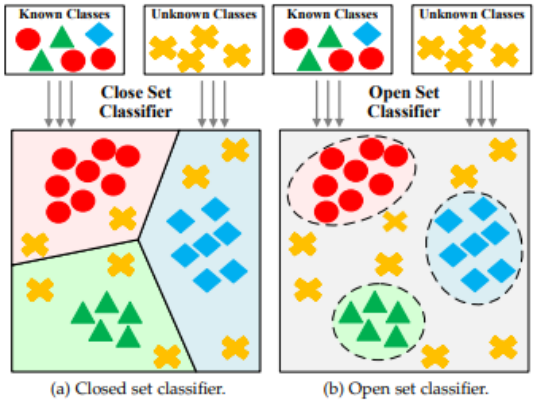

Figura 1: A diferença entre modelos a classificação closed set e open set [3]

que podem corresponder a comportamentos maliciosos. Assim, reutiliza-se a arquitetura original do TrafficGPT, inicialmente treinada utilizando o Deep Content dataset [5], composto por tráfego criptografado oriundo de vídeos do YouTube para entendimento do seu funcionamento . Para avaliar a capacidade do modelo de operar fora do domínio considerado pelos autores do artigo de referência [1], empregou-se um subconjunto representativo do CIC-IDS2017 [4], uma base amplamente utilizada na literatura por conter tráfego real misturado a diversos ataques simulados. Foram selecionados três arquivos: um contendo apenas tráfego benigno (segunda-feira), outro com ataques DoS (quarta-feira) e um terceiro com ataques de infiltração (quinta-feira). Os dados benignos e de DoS foram unificados para compor o subconjunto fechado utilizado nos testes, enquanto o conjunto de infiltração foi mantido separadamente para a simulação de ataque *zero-day*, permitindo verificar se o modelo seria capaz de rejeitar instâncias não vistas durante o treinamento. Nos experimentos realizados na MSI I, observou-se que o classificador K-LND rejeitou adequadamente essas instâncias desconhecidas, alcançando aproximadamente 65% de acurácia no cenário open set.

Contudo, a ideia central deste trabalho de Monografia em Sistemas de Informação II vai além da simples replicação do modelo: busca-se analisar o quão confiável o *TrafficGPT* realmente é quando confrontado com tráfego zero-day. Essa análise torna-se essencial porque LLMs, embora poderosos, são conhecidos por apresentarem fenômenos de alucinação [6], produzindo respostas altamente confiantes mesmo quando estão errados. Em um sistema de vigilância de tráfego, esse comportamento é potencialmente perigoso. A literatura recente demonstra que os modelos baseados na ativação do softmax, como o *TrafficGPT*, tendem a ser estruturalmente superconfiantes, especialmente em exemplos fora do domínio de treinamento [7]. Assim, investigar a confiabilidade do GPT-2 em um ambiente adversarial é tão relevante quanto avaliar sua acurácia.

Nesse sentido, o trabalho incorpora técnicas de calibração de confiança para mitigar a superconfiança natural do modelo. O método *Temperature Scaling*, proposto por Guo et al. [7], é aplicado como etapa de pós-processamento dos logits (pontuações internas atribuídas às classes e determinam a forma da distribuição de probabilidades), ajustando apenas a "temperatura", ou seja, aplica-se um parâmetro escalar que atua reescalando os logits antes da aplicação da função softmax. A otimização do parâmetro de temperatura é realizada via algoritmo L-BFGS-B [8], o que permite identificar se há superconfiança sem ter que executar todo o modelo base desde o treinamento. Por último, o trabalho discorre como os componentes da arquitetura do *TrafficGPT* conversam entre si, e o quanto alterar os valores brutos dos vetores, que são os objetos centrais para a classificação open set, impactam no desempenho dos classificadores como o K-LND, sendo este o classificador com melhores resultados no artigo de referência desse trabalho (Ginige et al.,2024) [1].

Este artigo está organizado da seguinte forma: a Seção II apresenta o referencial teórico e os trabalhos correlatos. A Seção III descreve a metodologia e as atividades conduzidas, incluindo o processo de investigação para o entendimento dos componentes que determinam a classificação e a confiança do modelo. A Seção IV discute os resultados, conclusões e sugestões para trabalhos futuros. Por fim, a Seção V apresenta as referências bibliográficas utilizadas.

II. Referencial Teórico

A crescente adoção de técnicas de aprendizado de máquina na área de segurança de redes tem estimulado a pesquisa em modelos que operem com conhecimento parcial ou incompleto sobre os dados em tempo de inferência. Nesse contexto, a abordagem de *Open Set Recognition* — estudada por Geng et al. [9], propõe mecanismos para que os modelos não apenas classifiquem corretamente amostras conhecidas, mas também consigam rejeitar ou identificar como "desconhecidas" aquelas que não pertencem ao conjunto de classes observadas durante o treinamento. Essa capacidade é especialmente relevante frente a ataques *zero-day*, nos quais os padrões de tráfego malicioso são inéditos.

Dentre os avanços recentes nessa área, destaca-se o modelo *TrafficGPT*, proposto por Ginige et al. [1], uma das primeiras arquiteturas a utilizar *Large Language Models* (LLMs), como o GPT-2, para tarefas de classificação de tráfego criptografado em contextos *open set*. A proposta visa explorar a habilidade dos LLMs em lidar com dados sequenciais, aplicando mecanismos de atenção para extrair representações contextualmente ricas a partir de sequências de pacotes. A arquitetura do *TrafficGPT* combina três etapas principais: (i) pré-processamento específico do dataset, (ii) extração de *features* contextuais via GPT-2, e (iii) classificação *open set*, utilizando técnicas como *K-LND*. Na avaliação original, o *TrafficGPT* foi testado com cinco conjuntos de dados de tráfego cifrado, incluindo o DC dataset, composto por amostras de tráfego de vídeos do YouTube e apesar de conter apenas dez classes (dez vídeos distintos), o DC demonstrou ser uma base adequada para avaliar a capacidade do modelo em aprender padrões discriminativos de tráfego criptografado. No artigo, os autores realizaram particionamento artificial do DC em 40% de classes conhecidas (*closed set*)

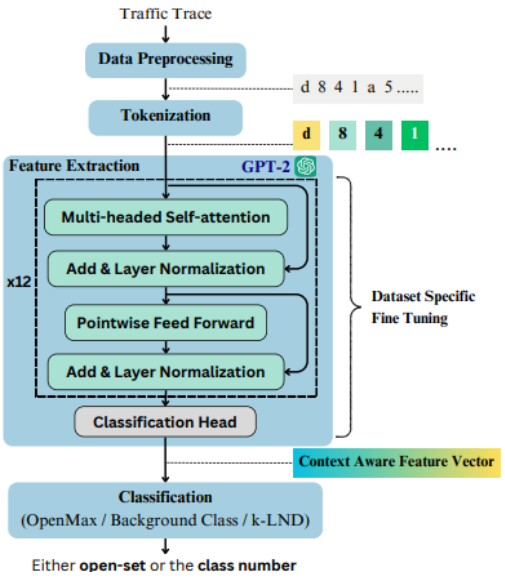

Figura 2: Arquitetura original do TrafficGPT proposta por Ginige et al. (2024), composta por três módulos principais: pré-processamento, extração contextual via GPT-2 e classificação open-set.

e 60% como classes inéditas (*open set*), permitindo avaliar a robustez do modelo frente a novos tipos de tráfego. Com pré-processamento adequado e extração de *features* via GPT-2, o *TrafficGPT* alcançou melhorias em relação a abordagens baseadas em outros modelos de redes neurais.

Entretanto, os dados utilizados no artigo original são, em sua maioria, limitados em variedade e escala, o que pode não refletir a complexidade do tráfego em redes reais. Na MSI I, testou-se o *TrafficGPT* frente ao dataset CIC-IDS2017, amplamente reconhecido na literatura de detecção de intrusos. O CIC-IDS2017 contém tráfego real e simulado, abrangendo uma gama diversa de ataques (DDoS, brute force, infiltration, entre outros) e atividades legítimas. Com suas amostras, pretendeu-se avaliar sua capacidade de identificar tráfego desconhecido e simular cenários de ataques *zero-day*. Essa abordagem segue o raciocínio de Ginige et al.. [1], que destacam a importância de modelos de detecção capazes de antecipar comportamentos maliciosos sem depender de assinaturas prévias. Assim, este trabalho explora o uso do *TrafficGPT* não apenas como um classificador tradicional, mas como uma ferramenta para reconhecimento de padrões fora do conjunto conhecido, um passo essencial em direção a sistemas de segurança proativos, resilientes e com maior capacidade de generalização. No entanto, para discutir e fundamentar o que seria a confiabilidade de um modelo em segurança de redes, é necessário reconhecer que redes neurais modernas, incluindo arquiteturas baseadas em Transformers como o GPT-2, apresentam uma tendência estrutural à superconfiança [6], produzindo previsões de alta probabilidade mesmo quando estão equivocadas ou expostas a dados completamente fora da distribuição demonstram de forma sistemática que modelos de aprendizado profundo contemporâneos, apesar de alcançarem excelente acurácia, são mal calibrados, isto é, suas probabilidades previstas não correspondem à verdadeira frequência de acerto [7].

Um estudo robusto e recente, "A Survey of Confidence Estimation and Calibration in Large Language Models", de Geng et al. [10] consolida a literatura mostrando que LLMs, incluindo GPT-2 e variantes modernas, mantêm altos níveis de confiança mesmo diante de entradas incertas, ambíguas ou fora do domínio, apresentando comportamentos como *overconfidence hallucinations* e incapacidade de expressar incerteza de forma alinhada à realidade. Esse fenômeno é ainda mais acentuado em modelos de aprendizado profundo como LLMs, cujas arquiteturas altamente parametrizadas tendem a gerar logits de grande magnitude. Tal característica induz distribuições altamente polarizadas após a aplicação da softmax, resultando em níveis de confiança artificialmente elevados, mesmo diante de entradas ambíguas ou fora do domínio de treinamento. Importante destacar que, embora os logits e as probabilidades produzidas pela softmax constituam representações vetoriais, o "nível de confiança" corresponde a um valor escalar, tipicamente a maior probabilidade atribuída entre as classes. Em outras palavras, apesar de os logits formarem um vetor de pontuações internas, a decisão final de confiança reduz-se à seleção de um único elemento desse vetor, correspondente à classe mais provável [10]. Assim, a superconfiança observada nos LLMs não reflete uma propriedade geométrica do vetor de saída, mas sim a concentração excessiva da distribuição após a softmax, um comportamento já discutido em estudos sobre calibração e incerteza, como apontado por Guo et al. [7]. Em contextos de segurança, essa característica é particularmente problemática: quando um modelo como o GPT-2 recebe tráfego zero-day (que ele jamais viu durante o treinamento), pode-se ainda assim "forçar" uma classificação conhecida com alta certeza, mascarando a incerteza natural do problema. Por essa razão, a confiabilidade deixa de ser apenas uma propriedade desejável e passa a ser um requisito essencial para o uso seguro de LLMs em sistemas críticos, pois somente um modelo calibrado é capaz de apoiar corretamente módulos como IDS, SIEM e classificadores open-set como o k-LND.

Assim, a adoção de técnicas de calibração como o *Temperature Scaling* torna-se uma possível estratégia para adequar o GPT-2, que tende a tomar decisões superconfiantes e, por isso, pode admitir comportamentos anômalos como se fossem legítimos, favorecendo a aceitação equivocada de tráfego malicioso. O *Temperature Scaling*, proposto por Guo et al. em 2017 [7], é uma técnica de pós-processamento aplicada diretamente aos logits produzidos pelo modelo neural, em que um único parâmetro escalar real positivo é aplicado a todos os logits e ajustado com base no conjunto de validação, refletindo o nível adequado de incerteza para as classes presentes.

No artigo [7], o valor da temperatura T é ajustado minimizando a *Negative Log-Likelihood (NLL)* usando o otimizador BFGS [13] padrão do *PyTorch*, sem impor limites ao parâmetro. Embora essa abordagem funcione, ela não impede que T assuma valores negativos ou numericamente instáveis, o que é problemático porque a temperatura deve ser sempre positiva

na fórmula da *softmax* "calibrada". Por isso, neste trabalho, foi

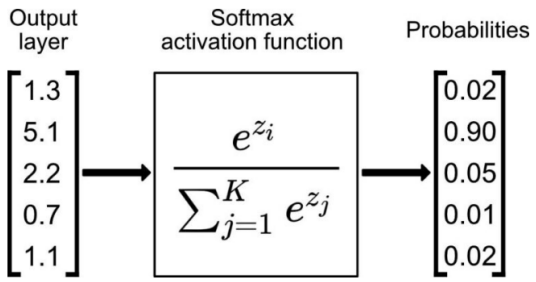

Figura 3: Representação do funcionamento da função softmax, que transforma os logits brutos produzidos pelo modelo em probabilidades normalizadas

utilizado o algoritmo *L-BFGS-B* que é uma versão aprimorada do BFGS que permite impor limites e garantir que *T>0* Além disso, o L-BFGS-B é um método clássico de otimização quasi-Newton para problemas suaves e de baixa dimensionalidade, ideal para ajustar o único parâmetro do *Temperature Scaling*. A formulação original desse método foi apresentada por Zhu et al. (1997) [8], mas a principal referência para a escolha dessa abordagem aplicada a uma LLM foi o artigo "Beyond Consensus: Mitigating the Agreeableness Bias in LLM Judge Evaluations"(Jain et al., 2025) [11]

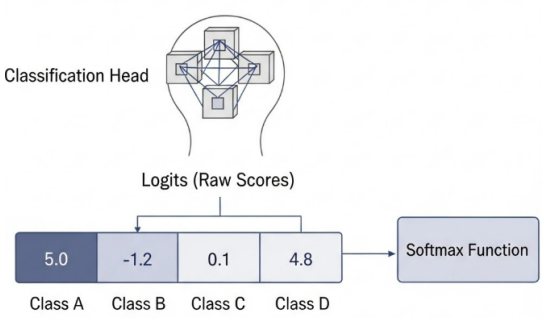

Figura 4: Importância de dividir logits por um parâmetro de temperatura T, realizando Temperature Scaling, que suaviza a distribuição e reduz a confiança

III. Contribuição do Trabalho

Este trabalho contribui em continuaro processo de adaptação do modelo *TrafficGPT* para um cenário simulado de detecção de tráfego anômalo, com foco na identificação de ataques zero-day, de modo a entender como a arquitetura do GPT-2 toma a decisão e é confiável, e qual seu impacto na tomada de decisão de admitir que um pacote de rede, por exemplo, seja admito na rede, ou seja, classificado como conhecido ou rejeitado.

A. Reprodução do TrafficGPT e treinamento do modelo

Na MSI I, o trabalho concentrou-se primeiramente em reproduzir o modelo *TrafficGPT* igual ao artigo [1], com o Deep Content Dataset [5]. E, posteriormente aplicá-lo a um cenário controlado de detecção de tráfego anômalo. O modelo foi treinado exclusivamente com o conjunto o CIC-IDS2017 [4] com dados de tráfego do subconjunto 'segunda-feira' (benigno) e 'quarta-feira'(ataque do tipo DoS); enquanto um segundo subconjunto,'quinta-feira' (ataque do Infiltração) foi utilizado apenas na etapa de avaliação para simular o contexto de zero-day. Esse processo permitiu analisar como o modelo se comporta ao classificar instâncias conhecidas e como reage quando exposto a padrões completamente novos, servindo como base para as investigações desta segunda etapa do projeto.

B. Redução da superconfiança do GPT-2

Durante o desenvolvimento desta segunda etapa do projeto, tornou-se necessário aprofundar o estudo sobre a origem da superconfiança em modelos de deep learning e, especialmente, em modelos baseados na arquitetura Transformer, como o GPT-2. Esse processo envolveu a leitura e análise de artigos fundamentais sobre calibração, confiança e comportamento probabilístico de redes neurais modernas, entre eles Guo et al. (2017), Geng et al. (2024), Jain et al. (2025) e trabalhos relacionados à formulação matemática do L-BFGS-B. Os estudos revelaram que a superconfiança não é uma aleatoridade, mas uma característica estrutural. Portanto, o *TrafficGPT* herda esse comportamento de alucinação. Porém, no artigo em que seu modelo foi proposto não há uma discussão acerca disso.

A literatura demonstra [7] [6] [10] que a função *softmax*, embora essencial para converter logits em probabilidades, tende a ampliar artificialmente as diferenças entre valores de ativação. Mesmo pequenas diferenças nos logits brutos podem resultar em probabilidades próximas de 1 após a exponenciação, criando a sensação de certeza absoluta. Assim, modelos como o GPT-2, treinados em larga escala e com alta capacidade expressiva, naturalmente produzem logits com magnitudes elevadas, o que aumenta ainda mais a polarização das saídas. Esse fenômeno explica por que modelos generativos podem classificar incorretamente (alucinar), devido à polarização excessiva da softmax e à magnitude elevada dos logits, o que atribui confiança máxima a respostas cuja plausibilidade estatística foi superestimada, mesmo quando os dados são ambíguos ou completamente fora do domínio.

Assim, avaliaram-se quais técnicas existentes poderiam reduzir esse comportamento superconfiante. A literatura aponta desde ajustes arquiteturais até métodos de pós-processamento. Entre as alternativas, a técnica *Temperature Scaling* destacou-se não apenas pela simplicidade, mas pela eficácia comprovada em LLMs. A técnica atua reescalando os logits antes da softmax, suavizando sua distribuição e corrigindo a discrepância entre confiança prevista e confiança real; para aplicar esse método corretamente, seria necessário encontrar um valor ótimo de temperatura *T* que minimizasse a *NLL* no conjunto de validação, presente no código `evaluate.py` do *TrafficGPT*.

Destaca-se que eficiência computacional e baixo custo de processamento e memória era fundamental para a realização do trabalho, dada as dificuldades e demora da execução do *TrafficGPT* discutidas na MSI I. Logo, foi nesse ponto que a busca por um método de otimização adequado se

tornou indispensável. O artigo original de Guo et al. sugere o uso de LBFGS, mas trabalhos mais recentes mostram que esse método, sem restrições, pode resultar em valores inválidos para T, como números negativos. Após investigar diferentes abordagens mais recentes [11], [12], concluiu-se que o algoritmo L-BFGS-B seria o mais apropriado para esta aplicação. O problema envolve otimizar apenas um parâmetro, e o L-BFGS-B é extremamente eficiente para problemas de baixa dimensionalidade, convergindo rapidamente e sem necessidade de retreinamento do modelo.

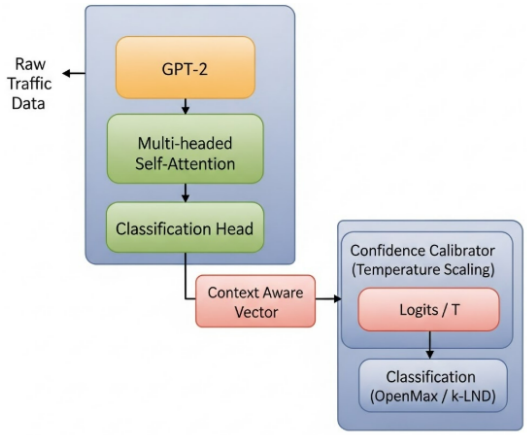

Figura 5: Representação da mudança arquitetural feita para obter logits menores que os originais

C. L-BFGS-B na otimização dos Logits

A aplicação do L-BFGS-B foi inicialmente realizada sobre o conjunto de validação do *Deep Content (DC)*, por ser um conjunto de dados menor, homogêneo e composto. Essa escolha permitiu testar a implementação de maneira mais simplificada, com maior controle sobre possíveis erros, além de reduzir significativamente o custo computacional nas primeiras iterações. Como o DC é o mesmo conjunto utilizado no treinamento do*TrafficGPT* e o modelo já havia sido ajustado por *fine-tuning* especificamente para esse domínio, era esperado que o comportamento dos logits fosse estável e que o modelo atribuísse probabilidades altas às classes corretas. A execução do L-BFGS-B confirmou essa hipótese ao retornar um valor de temperatura $T = 1.0$, indicando que não havia necessidade de recalibração: o modelo já se encontrava consistente entre treino e validação, sem sinais de superconfiança excessiva dentro desse domínio fechado.

Entretanto, ao repetir o mesmo processo no conjunto de validação proveniente do *CIC-IDS2017*, o comportamento foi diferente. A execução do L-BFGS-B encontrou um valor ótimo de $T = 1.4$, o que evidencia uma mudança significativa na distribuição dos logits. Esse aumento mostra que o modelo enfrentou maior dificuldade ao mapear amostras da validação do *CIC-IDS2017* para as classes, revelando uma discrepância natural entre o domínio do treinamento (benigno e DoS) e os novos padrões de tráfego (Infiltração). A necessidade de um $T > 1$ significa que os logits brutos estavam mais polarizados e que o modelo atribuía probabilidades mais altas do que deveria, mesmo em instâncias que não se assemelhavam ao conjunto original. Assim, a calibração evidencia que, fora do domínio de treinamento , o GPT-2 tende a produzir classificações com confiança aumentada, reforçando a importância da etapa de *Temperature Scaling* na formulação de um componente de rede para amparar a infraestrutura nos cenários de ataques *zero-day*.

D. Acurácias após a aplicação do Temperature Scaling

Não foram observadas mudanças significativas na acurácia do classificador open-set, K-LND, quando comparado ao resultado da arquitetura original ao executar o CIC-IDS2017. A aplicação do Temperature Scaling, embora eficaz para demonstrar a necessidade de ajuste a confiança do modelo base, não alterou substancialmente o desempenho do classificador open-set. Esse comportamento sugere que o escalonamento de temperatura atua apenas sobre a distribuição probabilística gerada pelos logits.

Tabela I: Comparação de acurácia entre arquitetura original e modificada no cenário Open-Set

Experimento	**Original**	**Modificada**
Acurácia Open-Set (%)	$\approx 65\%$	$\approx 63\%$

A metodologia consistiu em aplicar o algoritmo L-BFGS-B ao módulo `evaluate.py` com o objetivo de encontrar o valor ótimo de temperatura T para o conjunto de validação. Uma vez obtido o valor de T, os logits do modelo foram reescalonados por meio da divisão uniforme por esse parâmetro, preservando a ordem relativa entre as classes. Após essa etapa, seguiu-se o fluxo padrão de execução do *TrafficGPT* até a fase classificação open set realizada pelo método K-LND. A avaliação foi feita sobre o subconjunto correspondente à quinta-feira (ataques de infiltração), que representa o conjunto open set e contém apenas tráfego não visto durante o treinamento. A acurácia resultante foi então comparada aos valores obtidos na MSI I, permitindo verificar o impacto da calibração sobre o desempenho do K-LND. Os parâmetros, ferramentas e pré-processamentos foram mantidos idênticos entre as execuções, assegurando a equivalência metodológica entre os experimentos.

E. O impacto da calibração no K-LND e possíveis explicações

Observou-se, a posteriori, que a calibração por *Temperature Scaling* não alterou de forma significativa o comportamento do K-LND (tabela 1), evidenciando uma baixa correlação entre a redução da superconfiança dos logits e as decisões geométricas tomadas pelo classificador. Essa constatação emergiu somente após a análise do resultado e indica que o impacto do escalonamento de temperatura sobre o espaço vetorial utilizado pelo K-LND é limitado. Caso o *TrafficGPT* venha a ser considerado como um componente real de um sistema de defesa, essa relação deverá ser investigada de maneira mais profunda, explorando não apenas a calibração da confiança, mas também a influência das representações internas, dos vetores médios e das métricas de distância sobre o processo de detecção *open-set*.

Essa estabilidade da acurácia pode ser explicada pela natureza do próprio K-LND, que não toma decisões a partir das probabilidades produzidas pela *softmax*, mas sim a partir das distâncias geométricas entre os vetores de saída do modelo. Em outras palavras, o K-LND opera sobre os *Mean Average Vectors* (MAVs) [14] [15] e sobre a proximidade entre logits calibrados e não calibrados, em vez de depender diretamente do valor absoluto de confiança atribuído pelo modelo. A técnica *Temperature Scaling* atua apenas reescalando os logits por um fator T, preservando a ordem relativa entre eles. Como os logits de todas as classes são divididos pelo mesmo escalar, a direção do vetor e suas relações espaciais permanecem essencialmente as mesmas.

Consequentemente, as distâncias euclidianas entre os pontos no espaço vetorial sofrem apenas uma transformação linear uniforme, o que faz com que as medidas de proximidade utilizadas pelo K-LND mudem muito pouco ou permaneçam proporcionais. Assim, embora o *Temperature Scaling* seja eficaz para ajustar o nível de confiança do modelo, ele não altera de maneira significativa a organização geométrica dos embeddings produzidos pelo GPT-2. Por isso, o impacto observado na etapa de classificação *open-set* foi reduzido, indicando que melhorias mais profundas exigiriam atuar não apenas nos logits brutos, mas nas representações internas e nos vetores médios utilizados pelo *K-LND*.

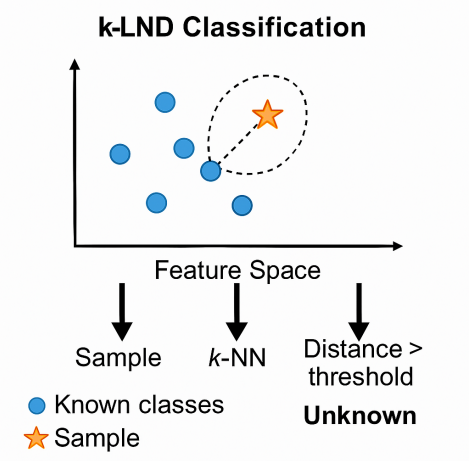

Figura 6: Demonstração de como o K-LND avalia a semelhança entre as classes

Conclusão e Trabalhos Futuros

O intuito original do modelo *TrafficGPT*, base para este trabalho, era propor um novo modo de classificar tráfego de rede criptografado interpretando-o como linguagem.Portanto, ele atende seu objetivo, mas para ser aplicado em um ambiente que requer proteção e resiliência cibernética, são requeridas melhorias como discutidas neste trabalho. Entre elas, destacam-se: (i) a necessidade de compreender em profundidade o nível de confiabilidade do GPT-2 e de seus mecanismos probabilísticos, dada sua tendência estrutural à superconfiança [7], [10]; (ii) a importância de calibrar suas saídas por meio de técnicas como Temperature Scaling, a fim de reduzir o risco de classificações incorretas com alta confiança; (iii) a investigação de como os logits e representações intermediárias influenciam o desempenho do classificador open set K-LND; e (iv) a necessidade de adaptar e aprimorar o modelo para que seja capaz de gerar vetores mais robustos, representativos e consistentes com a diversidade de dispositivos, padrões de tráfego e cenários presentes em redes reais.

Um dos achados relevantes desta investigação foi a constatação de que a calibração via técnica *Temperature Scaling* não se refletiu diretamente no desempenho do K-LND. Observou-se que existe uma baixa correlação entre a redução da confiança produzida pelo *softmax* e as decisões geométricas tomadas pelo K-LND, que opera com base em distâncias entre vetores e não nos valores probabilísticos calibrados. Essa diferença estrutural explica porque a acurácia do cenário *open-set* permaneceu praticamente inalterada após a calibração. O modelo arquitetural do *TrafficGPT* se torna mais honesto em termos de confiança, mas o espaço vetorial utilizado pelo modelo K-LND sofre apenas transformações lineares proporcionais. Ainda assim, o *Temperature Scaling* revelou-se útil para indicar quando o *TrafficGPT* encontra maior dificuldade em determinado domínio, funcionando como um sinalizador de que a arquitetura pode não estar adequadamente ajustada ao estilo daquele tráfego.

Dessa forma, embora este trabalho tenha contribuído para compreender e ajustar a confiabilidade do modelo, permanece aberto o desafio de melhorar o desempenho do modelo K-LND como classificador *open-set*. Avanços futuros devem considerar não apenas a calibração dos logits, mas também técnicas capazes de reorganizar o espaço latente, recalcular vetores médios ou adotar métricas alternativas de distância. Investigações mais profundas serão essenciais para que o *TrafficGPT* evolua de um modelo experimental promissor para um componente efetivo de defesa cibernética, capaz de operar com precisão e segurança em cenários reais de detecção de tráfego malicioso e, em especial, ataques *zero-day*.

Apêndice

As Figuras 4,5 e 6 deste trabalho foram geradas com o uso de inteligência artificial (ChatGPT com GPT-4) para ilustração de conceitos técnicos. As Figuras 1,2 e 3 são reproduções de diagramas originais dos artigos citados. As representações foram validadas com base na literatura técnica citada.

Referências

[1] Y. Ginige, Z. Zhan, and S. Wu, "TrafficGPT: An LLM Approach for Open-Set Encrypted Traffic Classification," in *Proceedings of the 2024 ACM Asia Conference on Computer and Communications Security (ASIA CCS '24)*, pp. 1–15, 2024. [Online]. Disponível: https://dl.acm.org/doi/10.1145/3674213.3674217

[2] L. Bilge and T. Dumitraş, "Before We Knew It: An Empirical Study of Zero-Day Attacks in the Real World," in *Proceedings of the ACM Conference on Computer and Communications Security*, 2012, pp. 833–844.

[3] Y. Wang, Y. Shi, Q. Qian, Y. Guo, Y. Chen, and M. Yang, "Orientational Distribution Learning With Hierarchical Spatial Attention for Open Set Recognition," *IEEE Transactions on Pattern Analysis*

and Machine Intelligence, vol. 45, pp. 8757–8772, Jul. 2023. doi: 10.1109/TPAMI.2022.3227913.

[4] I. Sharafaldin, A. Habibi Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," in *Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP)*, pp. 108–116, 2018. [Online]. Disponível: https://www.unb.ca/cic/datasets/ids-2017.html

[5] Z. Li and T. Huang, "Deep Content: Unveiling Video Streaming Content Using Encrypted Traffic," in *Proceedings of the 2016 ACM SIGCOMM Conference*, pp. 1–14, 2016. [Online]. Disponível: https://ieeexplore.ieee.org/document/8548317

[6] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," *ACM Transactions on Information Systems (TOIS)*, 2024. doi: 10.1145/3703155. [Online]. Disponível: https://doi.org/10.48550/arXiv.2311.05232

[7] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, vol. 70, pp. 1321–1330, 2017. [Online]. Disponível: https://arxiv.org/abs/1706.04599

[8] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-Scale Bound-Constrained Optimization," *ACM Transactions on Mathematical Software*, vol. 23, no. 4, pp. 550–560, Dec. 1997. [Online]. Disponível:https://doi.org/10.1145/279232.279236

[9] C. Geng, S. Huang, and S. Chen, "Recent Advances in Open Set Recognition: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3614–3631, Oct. 2021. [Online]. Disponível: https://doi.org/10.1109/TPAMI.2020.2981604

[10] J. Geng, F. Cai, Y. Wang, H. Koeppl, P. Nakov, and I. Gurevych, "A Survey of Confidence Estimation and Calibration in Large Language Models," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2024)*, pp. 6577–6595, 2024. [Online]. Disponível: https://doi.org/10.48550/arXiv.2311.08298

[11] S. Jain, U. Z. Ahmed, S. Sahai, and B. Leong, "Beyond Consensus: Mitigating the Agreeableness Bias in LLM Judge Evaluations," *arXiv preprint arXiv:2510.11822*, 2025. [Online]. Disponível: https://doi.org/10.48550/arXiv.2510.11822

[12] M. J. Zellinger and M. Thomson, "Rational Tuning of LLM Cascades via Probabilistic Modeling," *arXiv preprint arXiv:2402.05749*, 2024. [Online]. Disponível: https://doi.org/10.48550/arXiv.2501.09345

[13] D. C. Liu and J. Nocedal, "On the Limited Memory BFGS Method for Large Scale Optimization," *Mathematical Programming*, vol. 45, no. 1–3, pp. 503–528, 1989. [Online]. Disponível: https://doi.org/10.1007/BF01589116

[14] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward Open Set Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1757–1772, Jul. 2013. [Online]. Disponível: https://doi.org/10.1109/TPAMI.2012.256

[15] T. Dahanayaka, X. Li, Y. Wang, and M. K. J. Sharar, "Robust Open-Set Classification for Encrypted Traffic," *Computer Networks*, vol. 223, p. 109567, 2023. [Online]. Disponível: https://doi.org/10.1016/j.comnet.2023.109991

[16] SingleStore, "A Guide to Softmax Activation Function," SingleStore Blog, 2023. [Online]. Acesso em: 10 dez. 2023. Disponível: https://www.singlestore.com/blog/a-guide-to-softmax-activation-function/