Análise Linguística Forense com LLMs locais

OSINT e AI

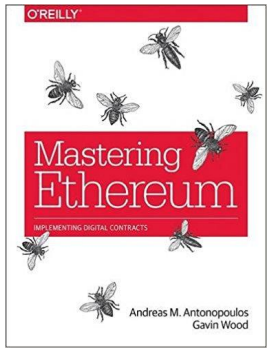
Everton Melo
por

VILLAGE OSINT

SOBRE

Pesquisador de segurança, entusiasta de projetos open source e sistemas descentralizados, contribui com mais de uma centena de projetos e produtos. Realizou diversas contribuições como organizador de eventos, possui vasta bagagem em infraestrutura, cloud, consultoria e inovação.

FreeBSD



minhas contribuições para o mundo.

Objetivo da palestra

Análise Linguística Forense.

LLMs (Grandes Modelos de Linguagem)

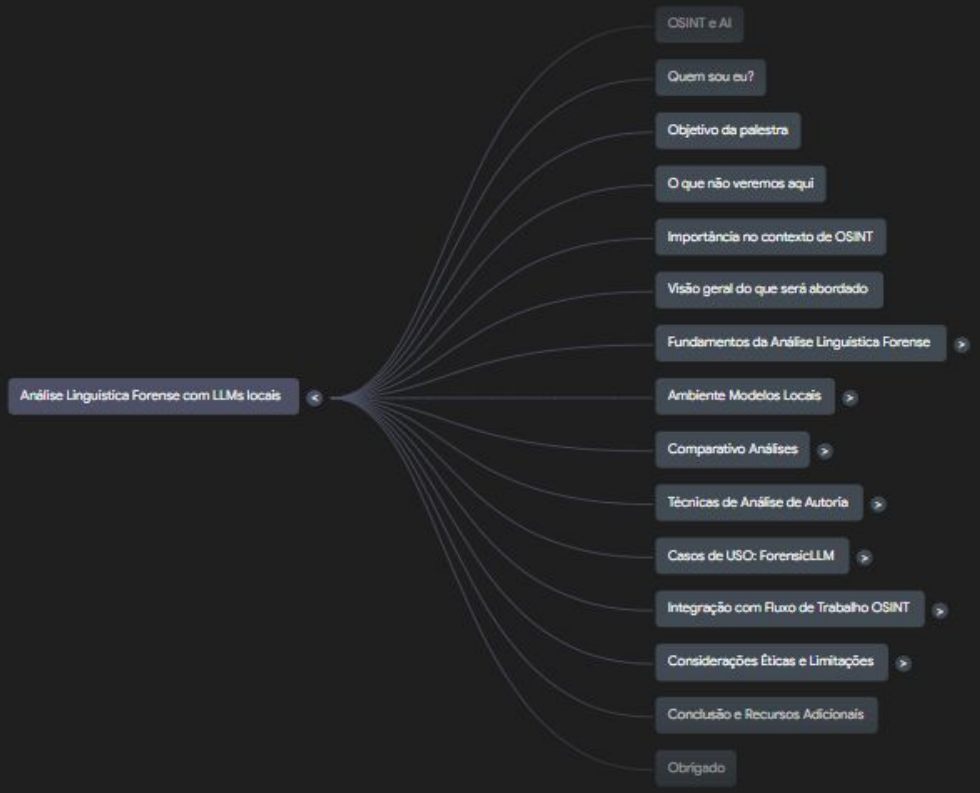
Estilometria.

Marcadores Linguísticos.

O que não veremos aqui.

RAG significa "Retrieval Augmented Generation" (Geração Aumentada por Recuperação).
Fine-Tuning: Fine-tuning é o processo de ajustar um modelo pré-treinado em um conjunto de dados adicional.
LGPD / GDPR / ISO
Documento com valor probatório.

Visão geral do que será abordado



Fundamentos da Análise Linguística Forense.

Fundamentos da Análise Linguística Forense

Conhecimento Linguístico

Aplicações de princípios linguísticos para resolver questões legais e investigativas.

Ciência Forense

Integração de técnicas forenses para análise detalhada de evidências.

Estatística

Uso de métodos estatísticos para identificar padrões em dados de comunicação.

Tecnologia de IA

Utilização de IA para detectar conteúdo gerado por IA em investigações digitais.

Estilometria Aplicada

A estilometria analisa quantitativamente o estilo linguístico para determinar autoria de textos. Baseada no princípio da "impressão digital linguística" única, utiliza algoritmos estatísticos para identificar padrões recorrentes como comprimento de frases, frequência de palavras e estruturas gramaticais preferenciais.
Na era digital, tornou-se essencial para identificar autores anônimos e diferenciar textos humanos de conteúdos gerados por IA.

Marcadores Linguísticos na Análise de Autoria

Marcadores Linguísticos

Elementos-chave para identificação de autoria

Escolhas Lexicais

Vocabulário individual usado por um autor

Estruturas Sintáticas

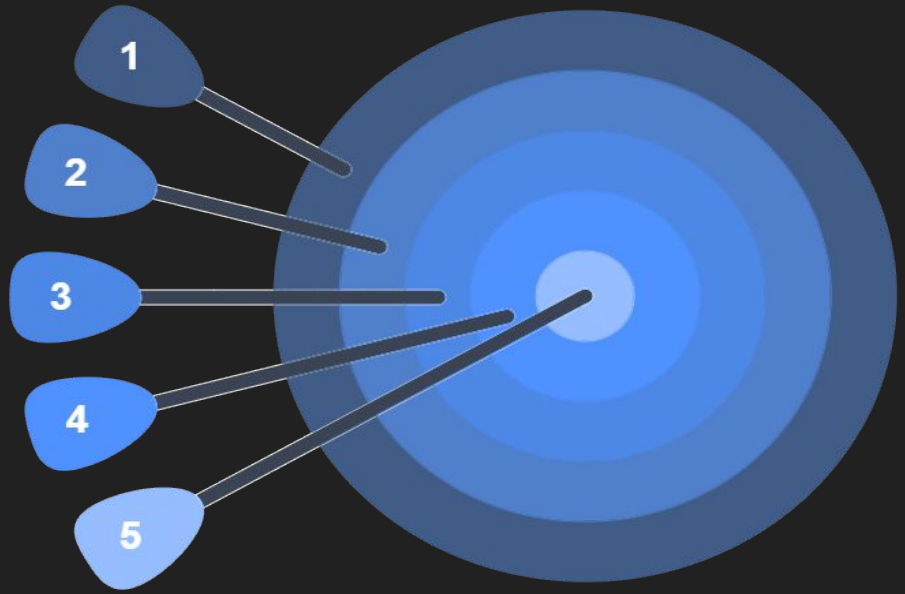
Organização frasal nos textos

Idioletos

Expressões características únicas de um autor

Estruturas Discursivas

Padrões organizacionais no texto



LLMs Processando Nuances

Os Modelos de Linguagem Grandes (LLMs) capturam padrões linguísticos complexos através de arquiteturas neurais avançadas. Utilizando mecanismos de atenção, identificam correlações entre elementos textuais, reconhecem ambiguidades semânticas e interpretam variações estilísticas. Desenvolvem representações vetoriais que codificam significados lexicais e estruturais, permitindo analisar marcadores linguísticos sutis como escolhas lexicais incomuns e construções sintáticas características.

Ambiente Modelos Locais

O ambiente local para análise linguística proporciona controle total sobre dados sensíveis, utilizando modelos como Llama, Mistral, MS Phi-* e R1. Llama destaca-se pela capacidade multilíngue, Mistral pela eficiência computacional, Phi pelo raciocínio lógico, e R1 pela consistência interpretativa. Configurado com hardware adequado, este ambiente permite executar análises sem dependência externa e oferece customização completa do pipeline analítico. A execução local facilita personalização para domínios específicos e adaptação a tarefas forenses especializadas, garantindo a documentação da cadeia de custódia em investigações.

LLMs suporte pt-br

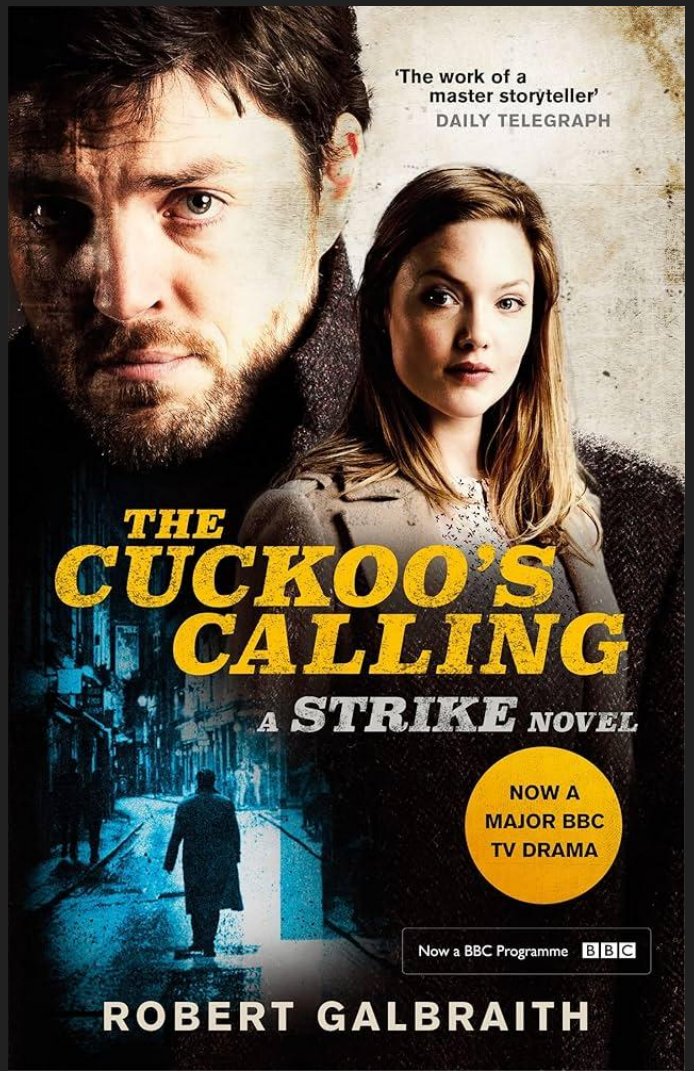
Família/Variante do Modelo	Parâmetros (Total/Ativo se MoE)	Comprimento do Contexto (Tokens)	Tamanho dos Dados de Treinamento (Tokens)	Foco Principal Declarado do Idioma	Suporte Explícito a Português	Idiomas Não Ingleses Suportados Chave (ou número)
DeepSeek R1	671B / (MoE)	128k	14.8T	Inglês, Chinês, Multilíngue	Não Especificado (foco em Eng/Ch)	Inglês, Chinês; outros fraco 25
Qwen 2.5 72B	72.7B	128k (até 131.072)	18T	Multilíngue	Sim	29+ incl. Ch, Es, Fr, De, It, Ru, Ja, Ko, Ar 1
Llama 3.1 70B	70.6B	128k	15T+	Multilíngue	Sim	De, Fr, It, Pt, Hi, Es, Th 6
Llama 3.3 70B	70B	128k	15T+	Multilíngue	Sim	De, Fr, It, Pt, Hi, Es, Th 2
Llama 4 400B (Maverick)	400B (Total) / 17B (Ativo)	1M	~22T	Multilíngue, Multimodal	Sim	Ar, De, Fr, Hi, Id, It, Es, Tl, Th, Vi 12
Microsoft Phi-3.5-mini	3.8B	128k	3.4T	Multilíngue	Sim	~23 incl. Ar, Ch, Fr, De, Es, Ru, Ja, Ko 9
Microsoft Phi-3.5-MoE	16x3.8B / 6.6B	128k	4.9T (10% multilíngue)	Multilíngue	Sim	~23 incl. Ar, Ch, Fr, De, Es, Ru, Ja, Ko 71
Microsoft Phi-4-mini	3.8B	128k	5T	Multilíngue	Sim	20+ incl. Ar, Ch, Fr, De, Es, Ru, Ja, Ko 3
Microsoft Phi-4-multimodal (texto)	5.6B (total)	128k	5T (texto)	Multilíngue, Multimodal	Sim (para texto)	20+ incl. Ar, Ch, Fr, De, Es, Ru, Ja, Ko 3

Comparativo Análises

A análise local oferece controle completo sobre dados, independência de conectividade, customização avançada e conformidade regulatória, embora exija maior investimento inicial. Já os serviços em nuvem proporcionam acesso imediato via APIs, escalabilidade instantânea e recursos computacionais superiores, mas apresentam limitações em investigações sensíveis: dependência de conectividade, exposição de dados confidenciais, custos variáveis e restrições de customização. Para investigações OSINT, a escolha depende da sensibilidade dos dados, recursos disponíveis e necessidade de controle sobre o processo analítico completo.

Mistério de: Robert Galbraith

Em abril de 2013, o romance policial The Cuckoo's Calling foi publicado no Reino Unido, creditado a um autor estreante chamado Robert Galbraith. A saga da revelação começou quando o jornal britânico The Sunday Times recebeu uma dica anônima via Twitter, em julho de 2013, sugerindo que J.K. Rowling era a verdadeira autora de The Cuckoo's Calling



Cuckoo's Calling e Estilometria

Os Softwares Utilizados:
Java Graphical Authorship Attribution Program (JGAAP) - Desenvolvido por Patrick Juola, este software é de código aberto e disponível gratuitamente.

Signature - Criado por Peter Millican, este sistema foi projetado para facilitar a análise estabilométrica com ênfase na identificação de autores

A metodologia incluiu a análise de diversos aspectos textuais:

- Comprimento de palavras, frases e parágrafos
- Frequência de palavras específicas
- Padrões de pontuação
- Uso de palavras funcionais (como "the", "to", "in")17
- Escolhas estilísticas específicas (como uso de frases em latim e pares adjacentes de palavras)

Professor Patrick Juola, da Universidade Duquesne (Pittsburgh, EUA)

Professor Peter Millican, da Universidade de Oxford (Reino Unido)

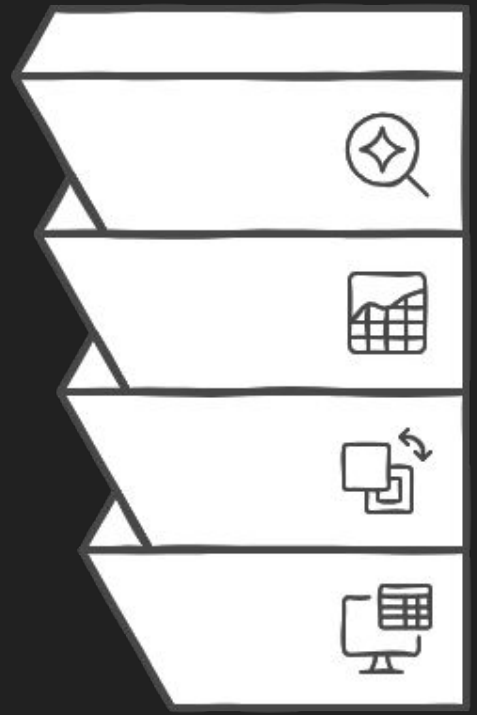
Técnicas de Análise de Autoria

Impressão Digital Linguística

A impressão digital linguística representa o conjunto único de características e padrões verbais que identificam um indivíduo como autor de textos. Similar à impressão digital física, é considerada única e difícil de falsificar conscientemente. Formada pela combinação de escolhas lexicais, estruturas sintáticas, hábitos de pontuação e organização do discurso, reflete influências educacionais, socioculturais e cognitivas do autor. Análises estilo-métricas computacionais conseguem quantificar estas características através de vetorização textual e modelagem estatística multidimensional, criando perfis linguísticos comparáveis. Na era das comunicações digitais e desinformação, tornou-se ferramenta crítica para atribuição de autoria em contextos investigativos e forenses.

Extração Marcadores Estilísticos

1. **Pré-processamento**
   Normalização e tokenização do texto
2. **Extração de Características**
   Identificação de características estatísticas
3. **Transformação Vetorial**
   Projeção de texto em espaço multidimensional
4. **Análise Automatizada**
   Uso de ferramentas para análise



Prompts Estabilométricos Especializados

Os prompts estilométricos especializados são instruções estruturadas que direcionam LLMs para identificar e analisar padrões linguísticos distintivos. Formulados com comandos precisos como "Identifique os 20 bigramas mais frequentes", "Extraia todas as estruturas sintáticas incomuns" ou "Quantifique o índice de subordinação frasal", estes prompts obtêm dados quantificáveis para comparação. A eficácia maximiza-se com instruções em camadas que primeiro estabelecem o contexto analítico, depois especificam métricas exatas e finalmente solicitam interpretação estatística. Prompts avançados incluem parâmetros para normalização de resultados, filtragem de elementos não-distintivos e visualização comparativa, transformando análises qualitativas em evidências quantitativas robustas para atribuição de autoria.

Metodologia Confiança Atribuição

A metodologia para quantificar confiança de atribuição combina abordagens estatísticas e linguísticas para estabelecer níveis probabilísticos de certeza autoral. Fundamenta-se em triangulação de múltiplas características independentes, análise discriminante multivariada e teste de hipóteses contra corpus de controle. O processo estabelece limiares estatísticos baseados em distribuições de referência, calculando índices como razão de verossimilhança (LR), que expressa probabilidade relativa entre hipóteses concorrentes. Protocolos rigorosos incluem validação cruzada, testes com autorias conhecidas e medidas para mitigar viases amostrais. A apresentação final incorpora intervalos de confiança e margens de erro, associando valores numéricos a descritores qualitativos padronizados (forte/moderada/fraca evidência) para interpretação forense contextualizada.

Precisão Modelos Comparada

A precisão comparativa entre modelos estilométricos varia significativamente conforme arquitetura e treinamento. GPT-4 lidera em análise contextual profunda com taxas de acerto de 87-92% em corpus controlados, enquanto modelos Llama-2 atingem 82-86% com melhor performance em textos técnicos específicos. Modelos menores como Mistral e Phi apresentam precisão de 76-80% mas com vantagens em eficiência computacional. Análises mostram correlação entre tamanho de modelo e performance, porém com rendimentos decrescentes após 13B parâmetros. Estudos comparativos indicam que especialização em domínios específicos frequentemente supera vantagens de escala, com modelos de 7B parâmetros ajustados para análise linguística forense superando modelos genéricos três vezes maiores.

Falsos Positivos Analisados

Os falsos positivos em atribuição estilométrica ocorrem quando textos de autores diferentes são incorretamente identificados como provenientes do mesmo autor. A incidência varia entre 8-15% em cenários investigativos reais, apresentando taxas mais elevadas em textos curtos (<500 palavras), conteúdo técnico padronizado e traduções. Fatores contribuintes incluem contaminação estilística entre colaboradores frequentes, adaptação intencional a estilos institucionais e convergência linguística em comunidades especializadas. Estratégias mitigadoras incluem estabelecimento de limiares probabilísticos conservadores, análise de métricas independentes, e incorporação de testes negativos com autores sabidamente diferentes. Modelos bayesianos avançados permitem quantificar e incorporar este risco na expressão final de confiança atributiva.

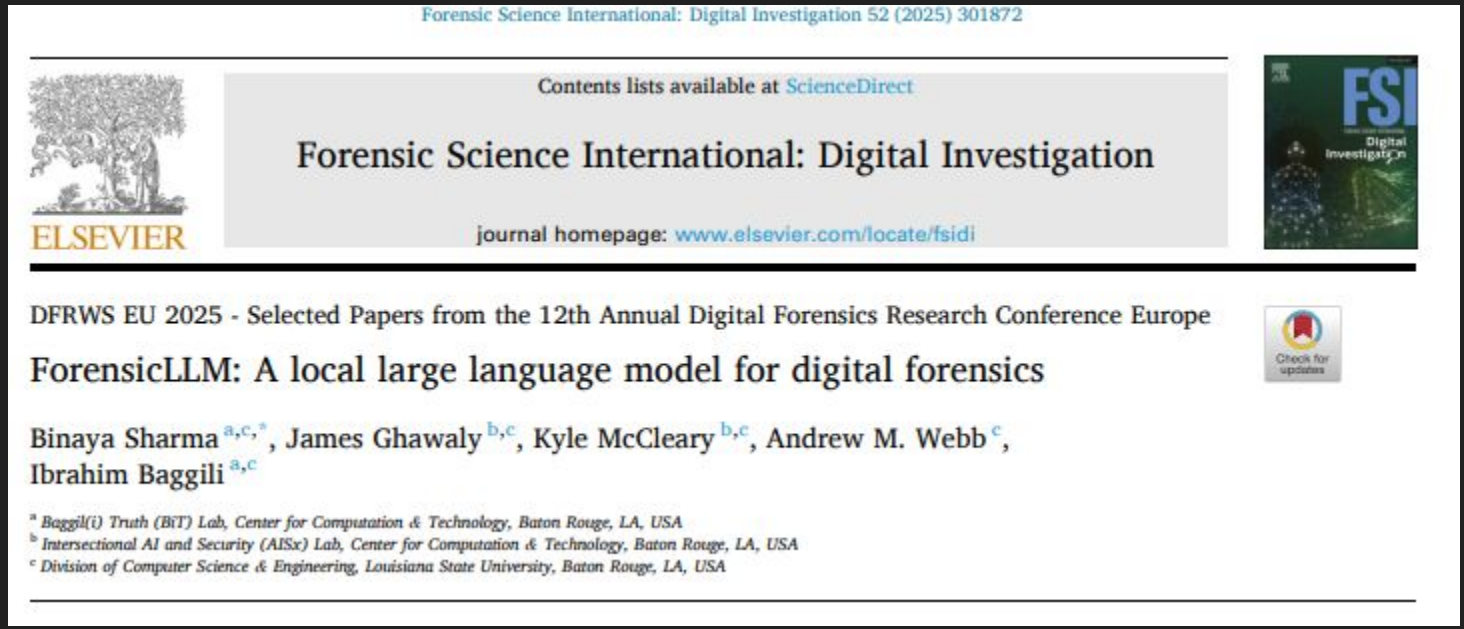
Limitações Técnicas Atuais

As limitações técnicas atuais em estilometria computacional compreendem obstáculos metodológicos e práticos significativos. Textos curtos (<300 palavras) frequentemente carecem de marcadores suficientes para atribuição confiável. Multilinguismo representa desafio particular, com desempenho reduzido em análises entre idiomas diferentes. Comunicações fragmentadas em plataformas diversas dificultam estabelecimento de corpus comparativo consistente. Adversários conhecedores podem empregar técnicas de ofuscação estilística como paráfrase algorítmica e mistura de estilos. Limitações matemáticas incluem dificuldade em estabelecer unicidade estatisticamente significativa em espaço estilístico superpopulado. Questões éticas e legais sobre privacidade e admissibilidade judicial como evidência independente permanecem controversas em muitas jurisdições. |

Casos de USO: ForensicLLM

o estudo de ForensicLLM

ForensicLLM é um Large Language Model (LLM) Local desenvolvido especificamente para a área de perícia digital (Digital Forensics - DF). Diferentemente de LLMs de propósito geral que muitas vezes dependem de APIs baseadas em nuvem ou computadores de alto desempenho e podem não ter especialização em campos como a perícia digital, o ForensicLLM foi criado para ser uma alternativa mais adequada



Forensic Science International: Digital Investigation 52 (2025) 301872

Contents lists available at ScienceDirect

Forensic Science International: Digital Investigation

journal homepage: www.elsevier.com/locate/fsidi

DFRWS EU 2025 - Selected Papers from the 12th Annual Digital Forensics Research Conference Europe

ForensicLLM: A local large language model for digital forensics

Binaya Sharma [a,c,*], James Ghawaly [b,c], Kyle McCleary [b,c], Andrew M. Webb [c], Ibrahim Baggili [a,c]

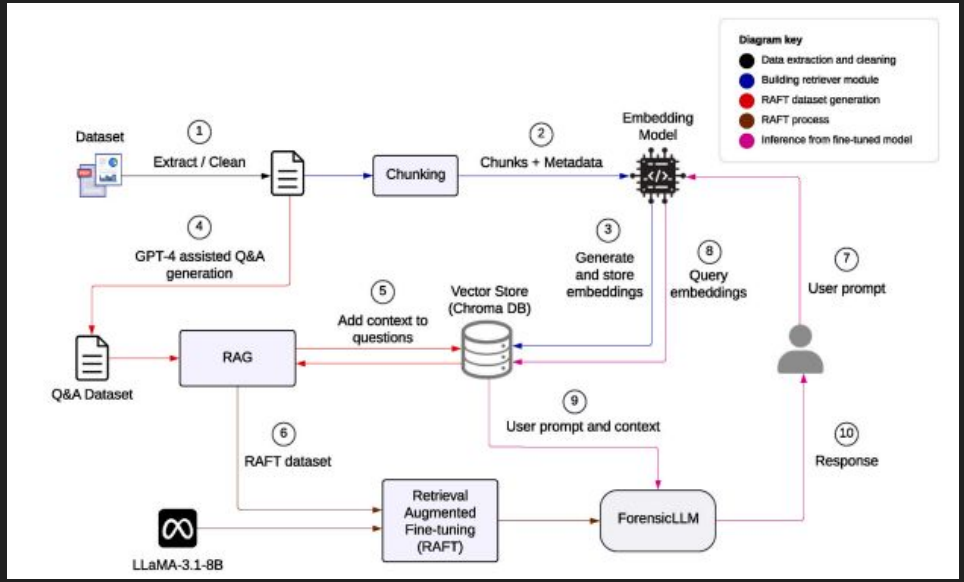
[a] Baggil(i) Truth (BiT) Lab, Center for Computation & Technology, Baton Rouge, LA, USA
[b] Intersectional AI and Security (AISx) Lab, Center for Computation & Technology, Baton Rouge, LA, USA
[c] Division of Computer Science & Engineering, Louisiana State University, Baton Rouge, LA, USA

Base Model: É um modelo LLaMA-3.1–8B quantizado em 4 bits. O LLaMA-3.1–8B foi escolhido por seu desempenho superior em comparação com modelos de tamanho similar e, com a quantização, pode rodar em GPUs de nível de consumidor, como a Nvidia RTX 4090.•

Treinamento e Otimização: Foi ajustado (fine-tuned) em amostras de Perguntas e Respostas (Q&A) extraídas de artigos de pesquisa em perícia digital e artefatos digitais curados.

DATASET ForensicLLM

1) Conjunto de Dados: O conjunto de dados para fine-tuning incluiu 1082 artigos de pesquisa em perícia digital da revista "Forensic Science International: Digital Investigation" e seu predecessor. Também utilizou dados de 1390 2)artefatos digitais do Artifact Genome Project (AGP), um repositório curado de artefatos forenses que passa por um procedimento de revisão para garantir validade e autenticidade



Fluxo de ForensicLLM

3) O AGP é um repositório de artefatos forenses que passa por um processo de revisão. Metadados de artefatos do AGP, como Título, Tipo, Dispositivo, Caminho, Descrição, Comentários, Search Tags e Dados, foram incluídos. Extração e Limpeza de Dados: Os dados dos artigos e artefatos foram extraídos e limpos.

4) Construção do Módulo de Recuperação: Foi construído um módulo de recuperação. Partes dos dados processados foram transformadas em representações vetoriais

Geração do Conjunto de Dados RAFT

5)Geração do Conjunto de Dados RAFT: Para realizar o fine-tuning com RAFT, foi necessário um conjunto de dados rotulado. Como não existia um conjunto de dados prévio em perícia digital para esse fim, o GPT-4 foi utilizado para criar o dataset. Usando um prompt cuidadosamente elaborado, o GPT-4 gerou pares de perguntas e respostas (Q&A) a partir dos artigos de pesquisa extraídos.
"instruction": Instruções para o modelo.

"input": Chunks recuperados do Chroma DB (o contexto) e a pergunta.◦

"output": A resposta desejada e as informações da fonte (título/autores).

Este dataset final foi dividido em conjuntos de treinamento (75%) e teste (25%).

o GPT na Criação do modelo?

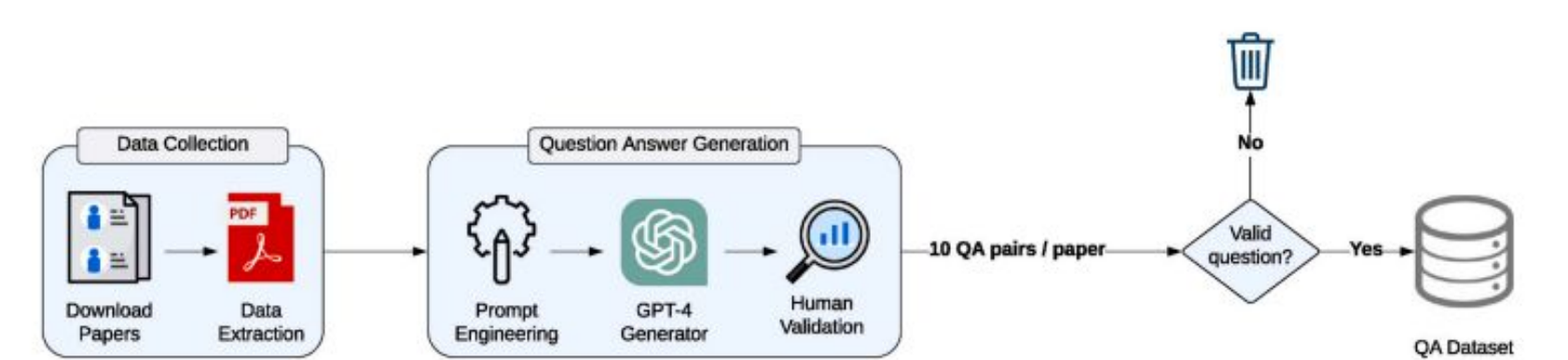


Fig. 3. GPT-4 assisted Q&A dataset creation process using DF research papers.

Capacidades do GPT-4 para Geração de Dados: A geração de datasets sintéticos e aumentados usando LLMs é uma tendência crescente. O GPT-4 foi alavancado para construir o dataset a partir dos dados extraídos de artigos de pesquisa, sendo uma escolha adequada devido ao seu "desempenho semelhante ao humano em compreensão de linguagem e sumarização de texto"

Resultado da Pesquisa

Com base nas informações fornecidas pelas fontes, os resultados da pesquisa sobre o ForensicLLM foram apresentados e avaliados em diferentes aspectos: desempenho quantitativo em um conjunto de dados de teste.

ForensicLLM

Superou todos os modelos em todas as métricas de avaliação.

LLaMA-3.1–8B + RAG

Desempenho consistentemente melhor do que o modelo base.

LLaMA-3.1–8B

Modelo base com o desempenho mais baixo em todas as métricas.

Model	Method	Response Length (in tokens)	BERTScore			BGE-M3	G-Eval
			Precision	Recall	F1 Score	Cosine	Overall
LLaMa-3.1-8B	One-shot	128.94	0.8929	0.8817	0.8872	0.8623	2.3787
LLaMa-3.1-8B + RAG	One-shot	226.38	0.8841	0.9010	0.8923	0.8805	2.6329
ForensicLLM	One-shot	156.34	**0.9215**	**0.9250**	**0.9232**	**0.9091**	**2.7544**

(o modelo base recebe apenas a pergunta sem contexto adicional, enquanto os modelos RAG e ForensicLLM utilizam contexto recuperado durante a geração da resposta).
Desempenho em Testes Quantitativos: todas as métricas quantitativas utilizadas para avaliação no conjunto de dados de teste (BERTScore F1, BGE-M3 Cosine e G-Eval Overall). Isso indica que as respostas geradas pelo ForensicLLM foram mais semanticamente precisas e bem avaliadas em comparação com os outros modelos

Detecção de Inconsistências e Desinformação

Inconsistências em Comunicações

As inconsistências em comunicações escritas são discrepâncias em textos que afetam sua credibilidade. Elas podem ser lógicas, estilísticas ou factuais e ocorrem quando a narrativa diverge da realidade ou de afirmações anteriores no mesmo texto.

Em análises forenses, estas inconsistências são fortes indícios de fabricação deliberada, múltiplos autores não declarados ou tentativas de ocultar informações.

Tecnologias avançadas de Processamento de Linguagem Natural (PLN) ajudam a identificar e quantificar estas anomalias.

Contradições Internas Analisadas

Contradições internas são incompatibilidades lógicas ou factuais dentro do mesmo texto. Podem ser:

Diretas: Afirmações que se excluem mutuamente.

Indiretas: Incompatibilidades em pressupostos ou implicações.

São comuns em textos fabricados, onde o autor não consegue manter uma consistência mental completa sobre a narrativa construída.

A detecção automatizada emprega algoritmos de inferência textual e modelos Transformer avançados, que mapeiam relações lógicas entre proposições e identificam contradições, mesmo as mais sutis e distantes no texto.

Inconsistências Temporais Identificadas

Inconsistências temporais ocorrem quando a cronologia dos eventos descritos apresenta impossibilidades lógicas, anacronismos (eventos fora de seu tempo) ou sequências implausíveis.

Manifestam-se como referências incompatíveis a datas, durações improváveis entre eventos ou conhecimentos anacrônicos em relação ao momento narrativo.

A análise computacional extrai marcadores temporais (explícitos e implícitos), construindo grafos que representam sequências. Algoritmos validam essas estruturas quanto à consistência, revelando falhas, especialmente em depoimentos fabricados.

Mudanças Estilísticas Detectadas

Mudanças estilísticas são variações não naturais nos padrões linguísticos de um autor. Exemplos incluem alterações abruptas em complexidade sintática, densidade lexical, preferências vocabulares ou estruturas retóricas sem justificativa contextual.

Frequentemente, indicam autoria múltipla, intervenção editorial significativa ou tentativas deliberadas de imitar outro estilo.

A detecção utiliza análise estilométrica e modelos neurais para identificar pontos de ruptura estatisticamente significativos e quantificar o grau de dissimilaridade entre segmentos textuais.

LLMs Extraindo Cronologias

Os LLMs demonstram capacidade excepcional para extrair cronologias ao processar narrativas complexas. Eles identificam e correlacionam marcadores temporais explícitos (datas, horas) e implícitos (referências como "na manhã seguinte").

Modelos especializados inferem relações sequenciais mesmo quando não explicitamente declaradas, construindo representações estruturadas. A eficácia é aprimorada com prompts que instruem o modelo a extrair primeiro os indicadores temporais e depois estabelecer suas inter-relações.

Esta abordagem permite reconstruir sequências de eventos, revelando inconsistências temporais sutis que podem indicar fabricação informacional

Verificação da Coerência Narrativa

A verificação de coerência narrativa examina a consistência lógica e causal entre eventos relatados, avaliando se a sequência forma um todo internamente consistente.

LLMs especializados analisam relações causais, motivações e plausibilidade contextual. O processo pode decompor narrativas em relações de evento-causa-consequência, validando antecedentes e consequências. Técnicas como rastreamento de entidades (Entity Tracking) ajudam a detectar saltos lógicos.

Esta metodologia identifica fabricações onde conexões causais implausíveis são introduzidas para sustentar narrativas desinformativas.

Mudanças Discursivas Sutis Identificadas

Mudanças discursivas sutis referem-se a alterações graduais ou estratégicas nos padrões argumentativos, enquadramentos ideológicos ou posturas retóricas em um texto.

LLMs identificam estas transformações através de análise de framing (como conceitos são contextualizados) e rastreando a evolução da polaridade emocional, posicionamento epistêmico (certeza/dúvida) e distância autoral.

Esta capacidade é crucial para identificar técnicas sofisticadas de desinformação, como a redefinição gradual de conceitos para manipular a percepção.

Protocolo de Análise de Autenticidade

Um protocolo prático para análise de autenticidade estabelece um fluxo de trabalho sistemático, combinando técnicas linguísticas, estatísticas e contextuais.

Inicia com caracterização estilométrica base.

Segue com verificação multiangular: consistência terminológica, coerência narrativa, validação temporal e estabilidade estilística.

Implementa triangulação metodológica, comparando resultados de diferentes algoritmos.

Requer documentação rigorosa de cada etapa, garantindo rastreabilidade e reprodutibilidade.

Incorpora análise de contra-hipóteses para mitigar vieses.

Integração com Fluxo de Trabalho OSINT

A integração com o fluxo de trabalho OSINT harmoniza a detecção de inconsistências textuais com inteligência de fontes abertas, criando uma poderosa sinergia analítica.

Implementa verificação cruzada bidirecional: inconsistências textuais orientam buscas OSINT, e dados OSINT contextualizam e validam anomalias linguísticas.

Pode estabelecer pipelines automatizados onde a análise textual alimenta queries em APIs de OSINT, criando ciclos de feedback.

Considerações Éticas e Limitações

LLMs: Limites Forenses

LLMs podem gerar "alucinações" (informações falsas convincentes) e são suscetíveis a vieses presentes nos dados de treinamento. Sua interpretabilidade (o "como" chegam a uma conclusão) é limitada, dificultando a validação forense. Não possuem verdadeiro bom senso ou compreensão contextual profunda, o que exige cautela ao aplicar suas saídas como evidência direta. A consistência e a reprodutibilidade exata das respostas podem variar, impactando a confiabilidade em investigações rigorosas.

Privacidade Dados Importam

O uso de LLMs em OSINT frequentemente envolve o processamento de dados pessoais. É crucial garantir conformidade com leis de proteção de dados, como a LGPD no Brasil. Avalie a legalidade da coleta e do uso dos dados inseridos nos LLMs, especialmente informações sensíveis. Considere os riscos de vazamento ou uso indevido e a necessidade de anonimização ou pseudonimização. A transferência internacional de dados para plataformas de LLM também requer atenção legal específica.

Documentação Rigor Essencial

Para que achados de OSINT com LLMs sejam admissíveis e defensáveis, a documentação é vital. Registre detalhadamente as fontes de dados, os prompts exatos utilizados, as versões das ferramentas e modelos de LLM, e as datas das análises. Mantenha uma cadeia de custódia digital para os dados coletados e processados. Assegure que os processos sejam auditáveis e, sempre que possível, reprodutíveis, explicando claramente como as conclusões foram alcançadas a partir das saídas do LLM.

Uso Responsável Tecnologias

A aplicação de LLMs em OSINT levanta questões éticas significativas. Considere o impacto potencial sobre os indivíduos investigados, incluindo o risco de discriminação devido a vieses algorítmicos. Promova a transparência sobre o uso dessas tecnologias em investigações. A responsabilidade pelas interpretações e ações derivadas das análises de LLMs recai sobre o investigador. Evite o uso para vigilância excessiva, manipulação ou para fins que violem direitos fundamentais.

Conclusão e Recursos Adicionais

Conclusão

Oportunidades inovação, Estilometria e melhoria.

- Inovação.
- Segurança de info.
- Educação
- Mentoria:

forense.melo@protonmail.com

bit.ly/consultorMelo

OBRIGADO