HTTPS://CONFERENCE.HITB.ORG/HITBSECCONF2024BKK

Words Have Meaning!

Leveraging LLMs to Enhance Insider Threat Investigation Capabilities

Keggy the Keg

DFIR Investigator/Engineer/Keg

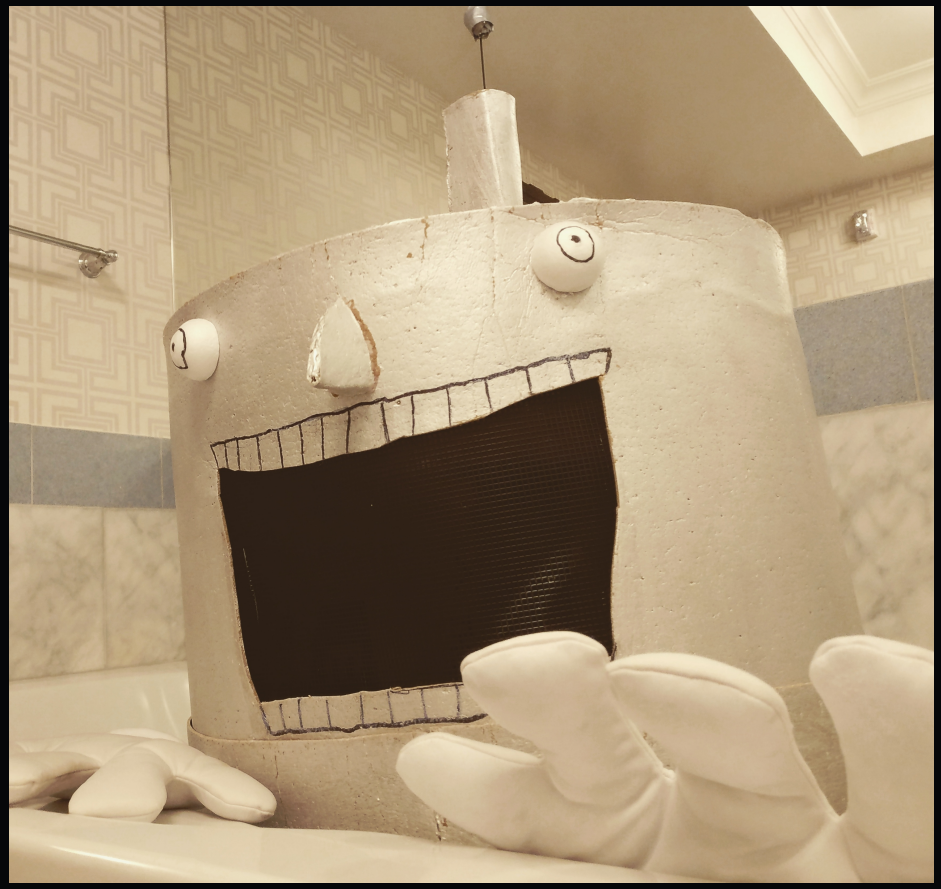
#HITB2024BKK

COMMSEC

30 AUG

Who is this?

- DFIR practitioner for 10+ yrs
- Investigator of cybersecurity/ forensics incidents
    - Everywhere from small businesses to [a major cloud provider]
- Recent focus on insider threats
- Ironically more of a scotch drinker



#HITB2024BKK

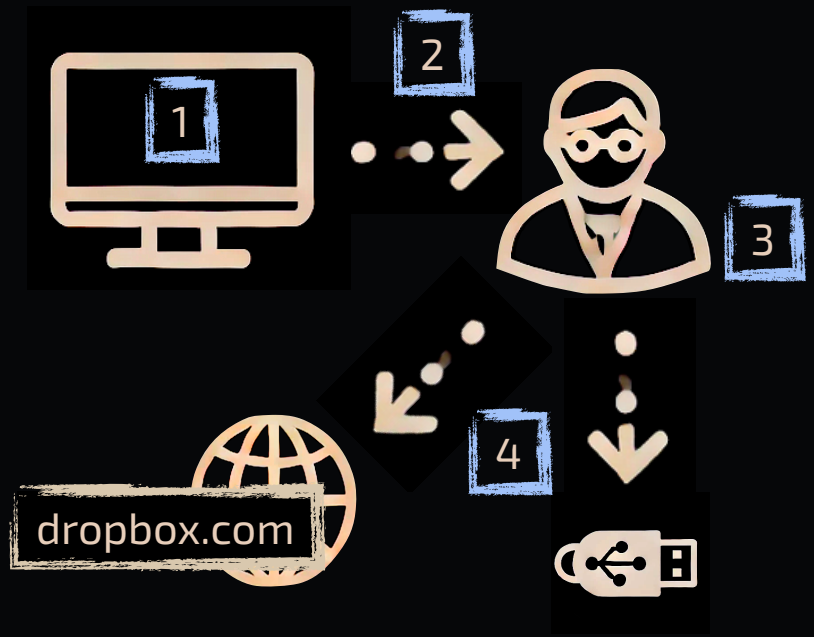
Agenda

- Background/Problem Space
- Solution 1 - Naive Method
- Solution 2 - Encoding Method
- Caveats/Limitations
- Open Questions & Future Research

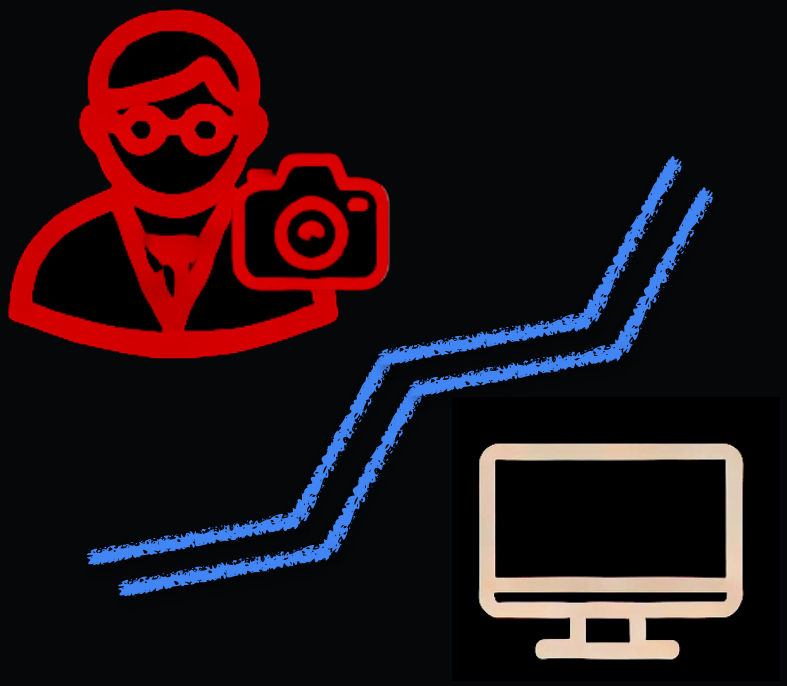
#HITB2024BKK

What's the problem?

- Goal of Digital Forensics is usually to tell a story
- Most leaks/insider threats follow a pattern:
    1. Data
    2. Access
    3. Retrieval/aggregation
    4. Exfiltration
- Many different motivations/actors
    - "Whistleblower"
    - Financial gain/IP theft
- If the entire trail is digital, high likelihood of successful attribution or alerting



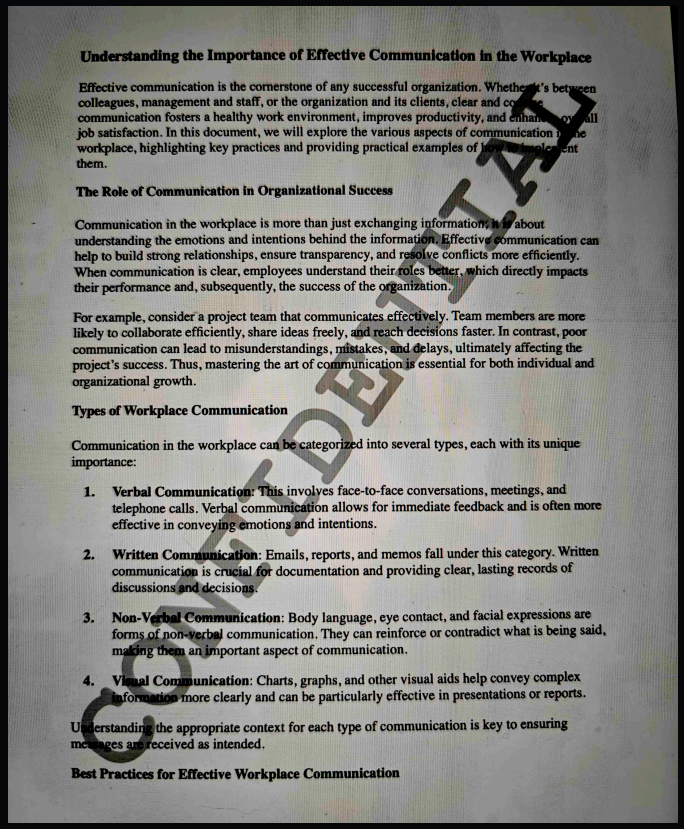
Mind the (air)gap



- A clever adversary may realize that they can break the digital trail by changing the medium of transmission
- Several ways to do this:
  - Take a picture of the screen
  - Read the document out loud
  - Allow someone to shoulder-surf
- Very frustrating to have to tell lawyers/clients/etc this could be a dead end
- But the *information content* still needs to remain intact for it to have any value
  - How do we trace information content to survive a change in medium?

Current State - DLP / Watermarks

- DLP is typically a suite of digital proactive and detective controls
  - Disabling/monitoring peripherals
  - Monitoring access patterns for aggregation or anomalies (see also UEBA)
- Watermarking can be digital or physical modifications to an item to enable some degree of traceability of its origin
  - Tesla use of spaces to encode user identifier
  - Other methods - font variations, page layout/alignment, etc
- All of these methods require either an unbroken digital trail or for the item to be transferred completely and intact



Understanding the Importance of Effective Communication in the Workplace

Effective communication is the cornerstone of any successful organization. Whether it's between colleagues, management and staff, or the organization and its clients, clear and co[illegible] communication fosters a healthy work environment, improves productivity, and enhance[illegible] overall job satisfaction. In this document, we will explore the various aspects of communication in the workplace, highlighting key practices and providing practical examples of how to implement them.

The Role of Communication in Organizational Success

Communication in the workplace is more than just exchanging information; it is about understanding the emotions and intentions behind the information. Effective communication can help to build strong relationships, ensure transparency, and resolve conflicts more efficiently. When communication is clear, employees understand their roles better, which directly impacts their performance and, subsequently, the success of the organization.

For example, consider a project team that communicates effectively. Team members are more likely to collaborate efficiently, share ideas freely, and reach decisions faster. In contrast, poor communication can lead to misunderstandings, mistakes, and delays, ultimately affecting the project's success. Thus, mastering the art of communication is essential for both individual and organizational growth.

Types of Workplace Communication

Communication in the workplace can be categorized into several types, each with its unique importance:

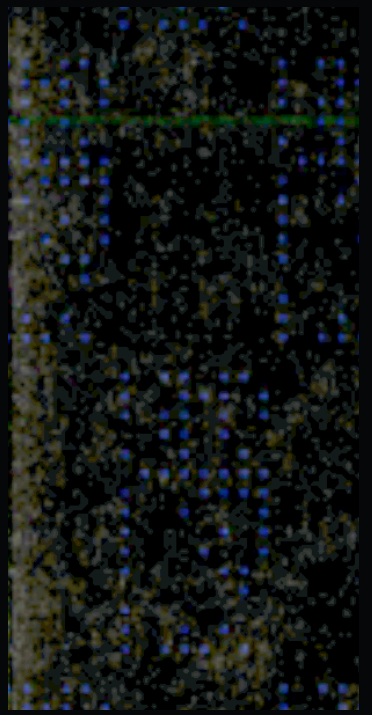
1. **Verbal Communication**: This involves face-to-face conversations, meetings, and telephone calls. Verbal communication allows for immediate feedback and is often more effective in conveying emotions and intentions.
2. **Written Communication**: Emails, reports, and memos fall under this category. Written communication is crucial for documentation and providing clear, lasting records of discussions and decisions.
3. **Non-Verbal Communication**: Body language, eye contact, and facial expressions are forms of non-verbal communication. They can reinforce or contradict what is being said, making them an important aspect of communication.
4. **Visual Communication**: Charts, graphs, and other visual aids help convey complex information more clearly and can be particularly effective in presentations or reports.

Understanding the appropriate context for each type of communication is key to ensuring messages are received as intended.

Best Practices for Effective Workplace Communication

Content Fragments

- If the entire document is recovered in the course of an investigation, attribution may be possible
  - Intercept & Reality Winner (she did not win)
- Challenge arises when only the most sensitive, controversial, important, or otherwise noteworthy portions surface
  - Can't rely on a single document-level watermark, need to identify highest likelihood targets and tag each
  - Good news - parsing for criticality to the overall document is a strength of LLMs



Source: https://archive.ph/g8FL9

Go Go Gadget LLMs

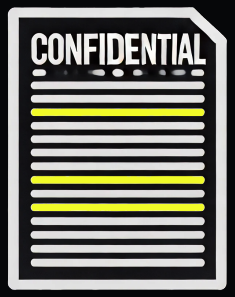
- In 2022, ChatGPT release begins current AI hype cycle - introduces public to concept of Large Language Models (LLMs)
    - Area of strength - parsing for semantic meaning
    - Creates output text that presentationally can pass for native speech
- Generation is stochastic (in default config)
- Non-deterministic nature an issue in certain areas
    - See "Mechanistic Interpretability" for efforts to resolve this
- What if we leverage these to watermark the most critical part of our information - the content?

#HITB2024BKK

Method #1 - Brute Force Overview

First method explores naive rewrite of sentences correlated to either individual users or target pools

1. Prompt LLM to identify most important sentences or semantically critical segments of the content



2. Prompt LLM to rewrite each segment. Tag each rewrite with a target user/group

User	Sentence
john.doe	[version_1]
jane.doe	[version_2]
jay.doe	[version_3]
jo.smith	[version_4]

3. Store rewritten segments and corresponding tags as entries in DB, one segment / table

4. In investigation, correlate identified segments to attribute or narrow suspects

#HITB2024BKK

Method #1 - Brute Force Prompt Variations

Phase 1:

"You are a linguistic expert specializing in the English language. Analyze the following document and respond with a JSON array of the most important sentences in the document."

Tested with:

- Senku 70B
- ChatGPT-4o
- LLaMA 3.1 8B, 70B
- Claude 3.5 Sonnet

Phase 2:

"You are a linguistic expert specializing in the English language. Rewrite the sentence [TARGET_SENTENCE] while maintaining the meaning of the original sentence."

vs

"You are a linguistic expert specializing in the English language. Rewrite the sentence [TARGET_SENTENCE] by changing the sentence as much as possible while still maintaining the meaning of the original sentence."

Tested with:

- Senku 70B
- ChatGPT-4o
- LLaMA 3.1 8B, 70B
- Claude 3.5 Sonnet

#HITB2024BKK

Method #1 - Brute Force Considerations

Drawbacks of this method:

❌ Combinatoric limitations*

- Only relatively limited number of ways to rewrite sentences, especially if they are not long
- Would likely not be able to scale to any reasonably sized org for individual sentences, have to settle for either narrowing target pool to 1/<number_of_permutations> or if separating by teams, could narrow pool to given team

❌ Storage Inefficiency

- Obviously requires more substantial storage for all permutations, and increases rapidly with scale of tool use (either of multiple sentences w/in doc or across many docs)

*NOTE: Could combine straight rewriting with other prompt features (“add one misspelled word”, etc) to expand possibility space a little bit

Method #2 - Encodings Overview

Second method explores targeted rewrite of fragments to maximize permutation space and create unique combinations of changes correlated to either individual users or target pools

1 Prompt LLM to identify most important sentences or semantically critical segments of the content

2 Prompt LLM to identify replacement candidates *within* each segment to create encoding space

```
[
{"pivotal": ["crucial", "critical",
  "key", "significant", "important"]},
{"moment": ["time", "point",
  "juncture"]}
]
```

3 Prompt LLM to rewrite each segment using a unique combination of the identified candidates. Map each rewrite as an encoding and associate it with a target user/group

User	Encoding
john.doe	0,1,4,2,3,1,0,5
jane.doe	1,1,3,2,3,0,3,5
jay.doe	3,2,0,3,1,1,3,3
jo.smith	1,2,1,3,4,1,2,2

4 Perform final LLM analysis of each rewritten encoding to ensure no unacceptable degradation or modification of content/meaning. Adjust or remove inappropriate encodings

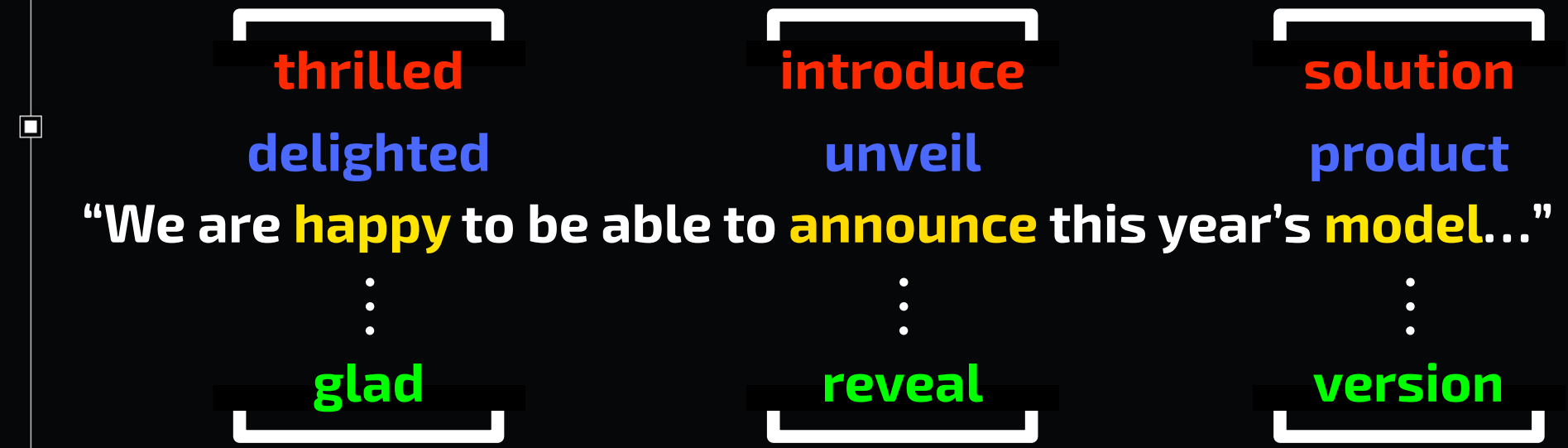
User	Encoding
~~john.doe~~	~~0,1,4,2,3,1,0,5~~
jane.doe	1,1,3,2,3,0,3,5
jay.doe	3,2,0,3,1,1,3,3

5 In investigation, can now correlate based on even *sub-segments of the encoding* to attribute or narrow suspects

#HITB2024BKK

Method #2 - Encodings Visualized

Here's an example:



#HITB2024BKK

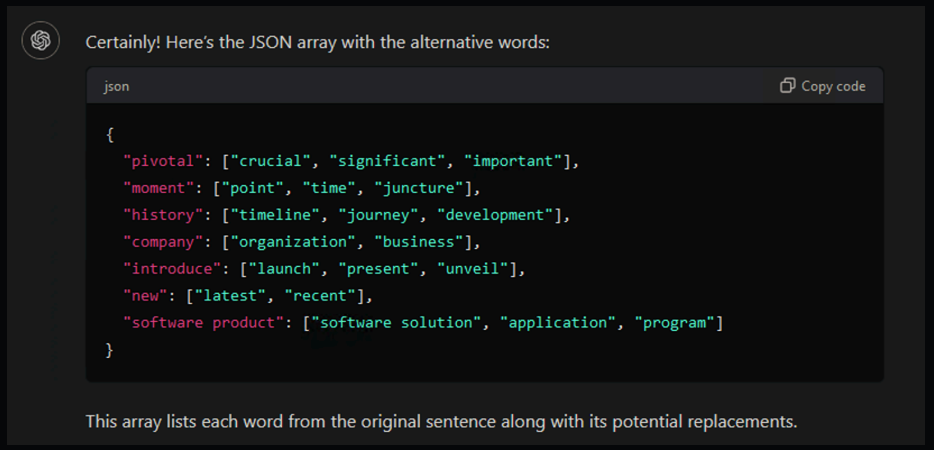
Method #2 - Encodings Prompt Choices

Phase 2 prompt:

> "You are a linguistic expert specializing in the English language. Analyze the sentence [TARGET_SENTENCE_HERE] and tell me which words could be changed without altering the meaning of the sentence. Your response should only consist of a JSON array mapping the original words to their alternatives."

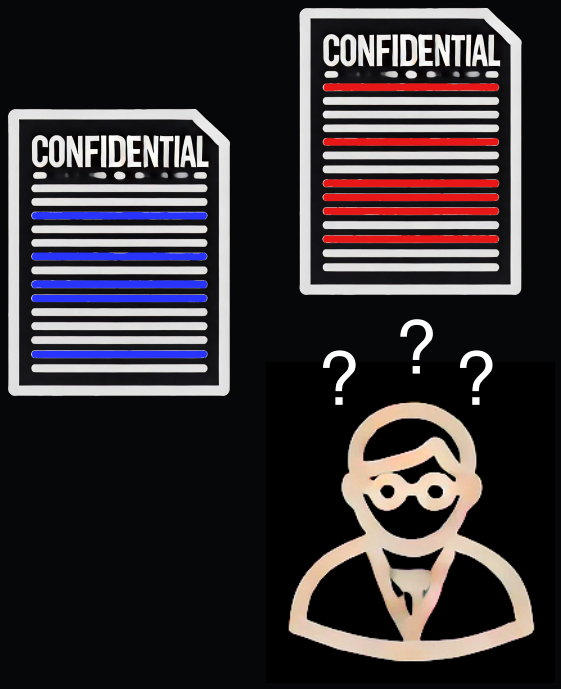
Tested with:

- Senku 70B
- ChatGPT-4o
- LLaMA 3.1 8B, 70B
- Claude 3.5 Sonnet



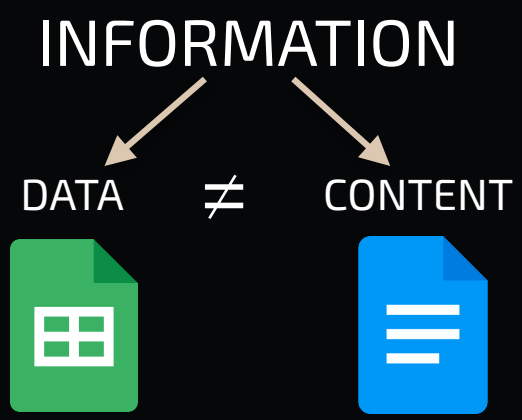
#HITB2024BKK

Caveats:
Comparing Variants



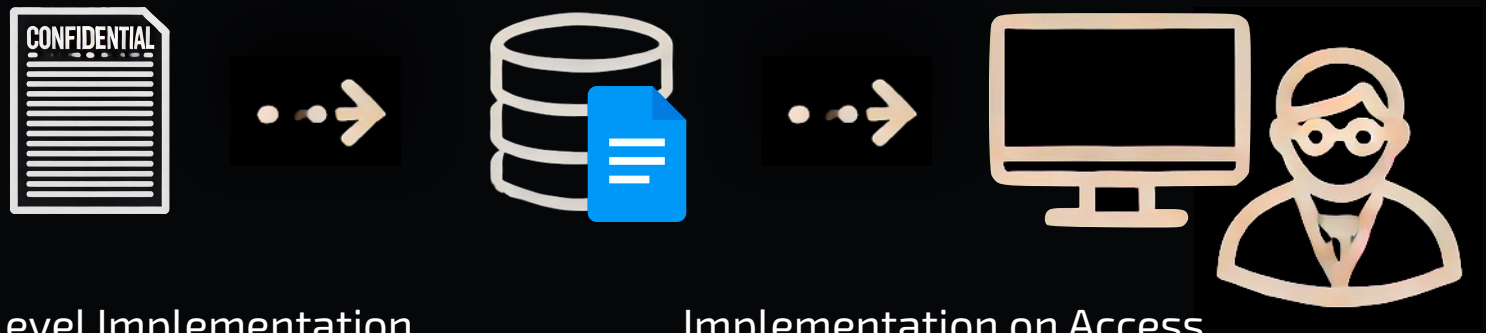
- Techniques described are vulnerable if one actor possesses multiple copies of the same content
  - Can happen innocently (error, screen sharing) or intentionally (seeking multiple sources)
  - Not unique to digital/semantic watermarking - physical barcodes/marks/etc also perceivable
- Identification of watermark can serve as deterrent or limit ability to share illicitly acquired information discreetly

Caveats:
Data vs Content



- Important to note: this method is only appropriate for when the **content** (aka semantic meaning) of the information is the key thing to protect
- If the structure/form of information must remain intact to be useful (i.e. data analysis, design specifications, engineering data, application code, etc) this technique is not a good idea
    - Example: For legally-privileged or sensitive documents, may not be able to modify document without legal implications
    - Could work with legal to review list of generated alternative wordings

Implementation Options



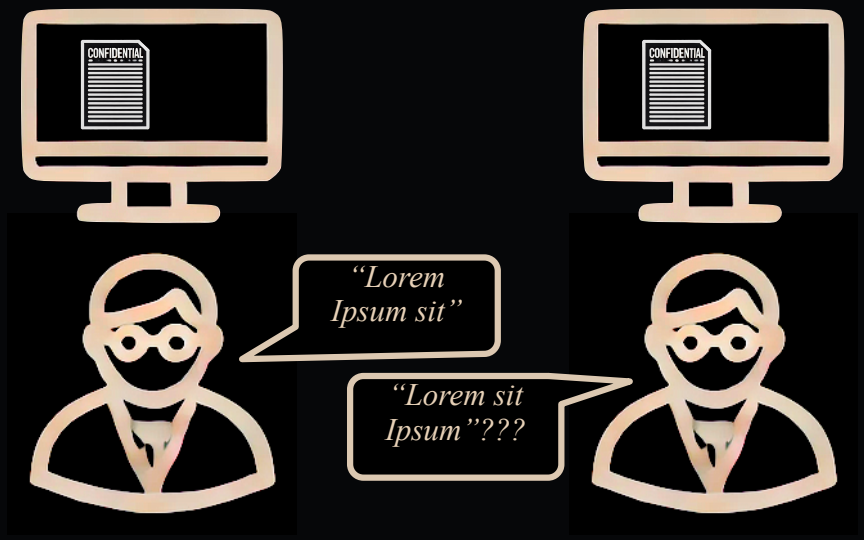
Application-level Implementation

- Integrate plugin into application housing/rendering the target content
- Enables pre-compilation of marked content - only computation required on access is assignment of variant to user/group
- More transparent to individual users
- More robust against efforts to circumvent marking of content

Implementation on Access

- Interposing between target data and end-user
- Examples:
  - Forcing access via a proxy to perform MITM-style injection of modifications to content
  - Leverage endpoint enterprise controls (enterprise-controlled browser extension, etc) to inject modifications
- More vulnerable to circumvention due to being client-side

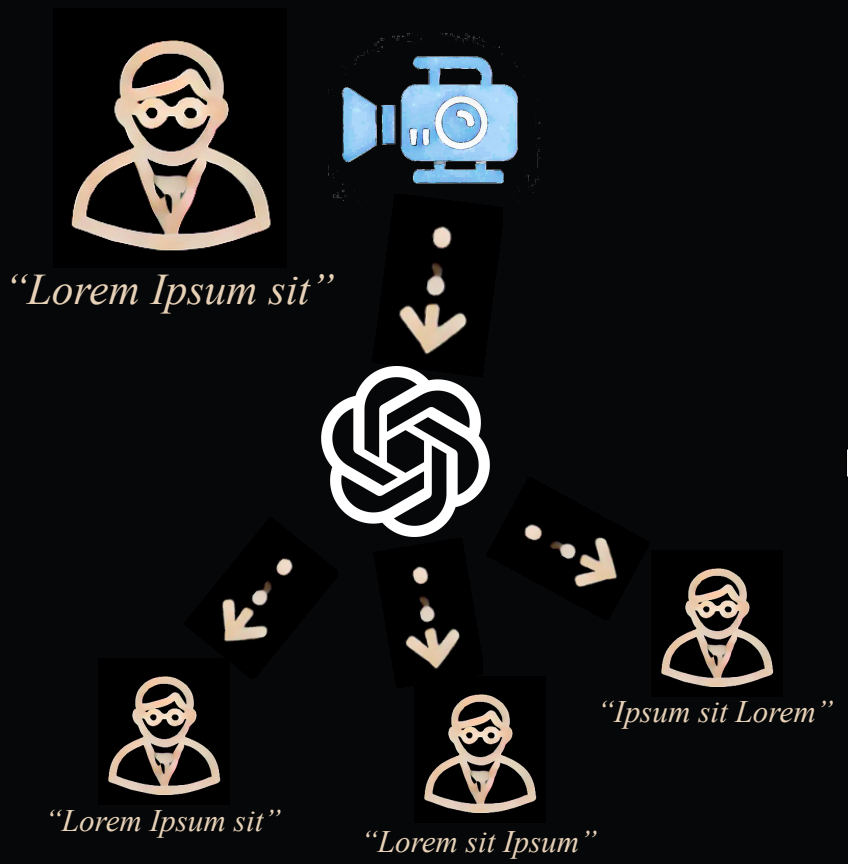
Open Questions



- Read-only nature of protected documents is a significant constraint
- Much writing work in modern enterprises is done via collaborative editing tools such as Office 365 or Google Docs
- Can't modify content when two or more users are simultaneously viewing same doc (and may be actively editing for meaning)

Beyond Text - Real-time Audio/Video

- Generative AI to modify audio/video is already widely available (Hugging Face, civitai, etc)
- Hardware requirements for real-time video and audio are currently demanding, but improved hardware and software likely to proliferate
- Potential further research on implementation of videoconferencing translation layer to analyze and modify content in near-realtime



#HITB2024BKK

Conclusions

Limitations

- Perceptibility - significant weakness of this technique
  - Reliant on only single instance, no comparisons
  - No collaboration
- Implementation - Some assembly required; no turnkey offerings
  - Can also be computationally/financially expensive based on number of versions required
  - Data privacy issues may arise if LLMs are not self-hosted

Strengths

- Technique's strengths are robustness and scalability
  - Ability to maintain watermark across transmission medium changes is highly robust
  - For small orgs with limited targets, not worth it probably
  - For large orgs with a rigorous document management system (oftentimes read-only by default) this could make sense
- Best use cases are broadcast mediums where information is sensitive (email, internal messaging)

Thanks! ขอบคุณมาก!

#HITB2024BKK