NER-D: ML Sensitive Data Detection

Find sensitive data by what it means, not what it looks like.

Legacy detection reads data by shape, one word at a time.

Two decades of data protection rest on one assumption: that sensitive data can be recognized by its shape. A credit card has sixteen digits; a Social Security number is three digits, two, then four. Catalog the shapes and scan for them. That held when sensitive data lived in structured fields and known documents. It cracks the moment the data is a concept rather than a format, and it breaks when the content is written live in an AI conversation, not prescanned and labeled.

The most valuable enterprise data has no fixed format at all: prescription drug names, trading strategies, or national identity numbers. Shape-based detection misses it, or false-positives on every pattern that resembles an identifier. It decays a little more each day the world changes. Prior NER moved past shape but hit a ceiling: to stay fast enough for live traffic it ran on small models that carry too little world knowledge to recognize a concept they were never trained on.

The name and the human test:

Paris, France
a place

Paris Hilton
a person

Plaster of paris
a material

Same letters, three meanings. Hard for legacy tools, natural for NER-D.

NER-D on the zero-shot benchmarks:

- **+10.8** F1 over the prior best method
- **20x** Faster than pubic models
- **20/20** Zero-shot benchmarks won
- **> 100** Languages supported

Catch what legacy tools miss

Protect the concept-level data that has no fixed format.

Formatless data, in scope: Prescription drugs, code secrets, and weapons are detected, not just credit cards and SSNs.
Meaning, not just shape: Detection reads what a string means in its context, the way a careful human reviewer would.
Real-time classification: Catches sensitive data written inside AI prompts, which no fingerprint or index has ever seen.

Cut false-positive noise

Significantly improve accuracy by eliminating 'lookalike' matches.

Context decides the call: The same nine digits read as a national ID in one sentence and an invoice total in the next.
No more alert floods: Context-aware classification end the alerts that fire on every number resembling an identifier.
Analyst time on real work: Lower false-positive volume means hours spent on real incidents instead of dismissing noise.

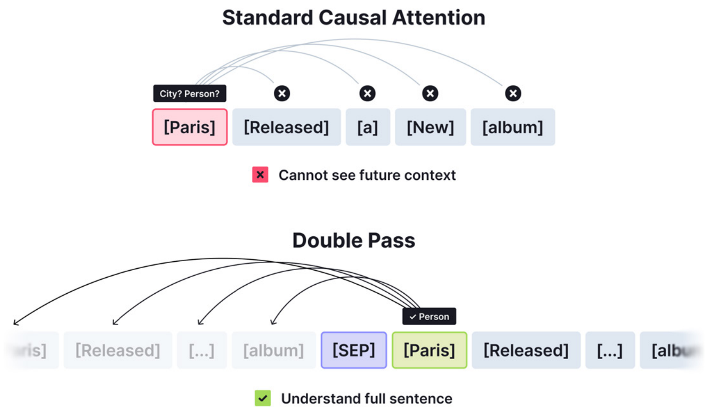
End the maintenance treadmill

Concept-level rules stay current without daily human tuning.

World knowledge built in: The model recognizes national ID formats, drug names, and trading concepts from its training.
No detector lists to build: Name a concept once and skip rebuilding lists each time the world adds 150 new drug names.
Add a type in a sentence: Define new sensitive data in plain English and deploy in seconds, with no labeled data or retraining.

How NER-D detects by meaning

- **Detection on a full LLM:** NER-D runs sensitive-data detection on a full large language model rather than a small, narrow one, so it draws on world knowledge from training instead of a hand-built list, and does it at the speed real-time inspection requires.
- **Double-pass classification:** The input passes through the model twice, so every token is classified with the whole sentence already in view. Classification happens directly in one parallel pass, not token by token, which is where the speed comes from: over 20x faster than comparable generative methods.
- **Contextual disambiguation:** Because it reads meaning, NER-D reads context. The same string classifies differently by what surrounds it, so Paris France is different from Paris Hilton or plaster of paris. That is what turns high recall into low false-positive noise.
- **Zero-retraining extensibility:** Add a new type of sensitive data by writing a plain-English definition, ready at inference time with no labeled data, no model retraining or fine-tuning, and no maintenance burden.
- **World-knowledge coverage:** Name a concept such as national ID numbers and the model already recognizes the variants across countries and languages, so one rule replaces the regex pile that once matched nearly every number on the planet.
- **Benchmark-leading accuracy:** NER-D posts the highest zero-shot F1 of any method benchmarked, +10.8 over the prior best, winning all 20 datasets of an extended suite. The method is peer-reviewed and accepted at ACL 2026, so the numbers are auditable, not marketing.
- **Additive to your DLP:** NER-D complements existing DLP rather than replacing it. Inside the WitnessAI platform it adds to the 100-plus data types already detected and feeds the workflows teams run today.



The double pass: world knowledge at real-time speed

Why an LLM can run inline.

Generative detection writes its answer one token at a time, which is slow and costly on live traffic. NER-D classifies every token in a single parallel pass, over 20x faster than generative methods at a fraction of the cost.

How the double pass works:

- **Read the whole sentence:** Passing the input twice lets every token see the full sentence, so later context can resolve meaning.
- **Classify in parallel:** Tokens are classified directly in one pass, not generated one at a time, creating speed.
- **Name a concept, not a list:** A rule names a concept and the model supplies the variants across markets and languages.
- **Refine in plain English:** When detection is off, edit the sentence. That is the entire tuning loop, with no retraining.

About WitnessAI

WitnessAI is the confidence layer for enterprise AI. Our AI-native platform provides network-level visibility and intent-based controls to govern your entire workforce, human employees and AI agents alike. We deliver runtime security for models, applications, and agents, deployable via network integration or lightweight APIs. Our single-tenant architecture ensures data sovereignty and compliance.

Learn more at witness.ai.