Beyond the Answer: Operationalizing AI on Classified Information

By: William Du, Founding ML Engineer and Kiam Boerema, Senior Account Executive

Classification Markings Govern Who Can Act on Information

Before an answer can be shared, someone has to establish how every piece of it must be handled.

Classification markings communicate how information must be protected and who may access or receive it. They are used across government, defense, intelligence, and other national-security environments wherever classified information is created, stored, processed, or shared. This information appears in reports, emails, briefings, and the digital systems that hold them.

A marked document carries several types of markings. Document-level markings describe the overall classification of a document, typically through banners at the top and bottom of each page. Portion markings identify the classification of specific paragraphs, bullets, tables, or other sections.

For example, a document may be marked **SECRET** overall while individual portions are marked **(U)** for Unclassified, **(C)** for Confidential, or **(S)** for Secret. Additional markings may specify handling or distribution restrictions, including access limited to particular compartments, programs, or authorized communities. Together, these markings allow users to understand both the document's overall protection requirements and the status of each part within it. They govern how that document should be handled and who can then see it.

An Answer That Cannot Be Shared Has No Operational Value

An assessment with the wrong classification level is a decision delayed or denied.

An intelligence product is only as valuable as an organization's ability to act on it. Even the most analytically-sound and valuable information is unusable if no one has yet reconciled where each piece came from and marked the product. The lack of classification fails organizations in three distinct ways, each with its own cost.

1. When classification is too slow, the answer misses its window; on a compressed timeline, a decision that arrives late is a decision not made.
2. When classification is too high, the answer cannot be shared with the people who need it. Or, it has to be pulled back and corrected, costing analysts even more time and effort.
3. When classification is too low, the result is spillage: information handled below its required protection level, with consequences that reach well beyond the mission at hand.

The cost of incorrectly classified information is inaction at best, but mission-compromising at its worst.

An AI-Generated Answer Without Markings Is Not Actionable

An agent can draft an assessment in seconds and leave the analyst an hour away from being able to send it.

AI agents and LLM workflows are extremely capable. They can read reports, find relevant passages, compare sources, and draft an assessment in seconds. While these outputs are useful, a user in a classified environment often needs more than a good answer. They need to know where their answer came from. Not only to trust the generation, but to also know how that answer must be handled.

AI platforms often have citations at a document or passage level that tell an analyst which source was retrieved and used to produce a generation. They add transparency to LLM generated text. However, these citations do not show how the answer must be handled. Classification banners and portion markings determine who can see the answer and how it can be shared. And oftentimes in these platforms, the work to attach a document or portion mark remains manual and reliant on the analyst. Unfortunately, when this happens, an AI platform has simply shifted manual work downstream.

In this case, the analyst is now responsible for reconstructing portion markings from the information the platform provides them. In practice, they must:

1. Understand which passages and documents contributed to the response
2. Determine what markings appeared in the source documents
3. Analyze which of the markings apply to generated content
4. Label the generated content with portion and document markings

In some cases, that process takes longer than drafting the entire product manually. We saw this during Scarlet Dragon 26-01, an exercise led by the Army's XVIII Airborne Corps. Analysts could not use AI-generated intelligence summaries that lacked classification markings because verifying the underlying sources and applying the correct markings took as long as writing the summaries by hand.

For an AI platform to deliver real value in a classified environment, classification markings must be integrated into every stage of the workflow, from information retrieval and answer generation to presentation and human review. An AI answer without markings is not actionable, however good the analysis behind it.

Classification Is a System-Level Responsibility

Reliable portion markings depend on preserving classification context across the entire AI pipeline.

An AI platform built for classified work must treat portion markings as a first class feature, the way citations already are. Generating reliable classification markings is not as simple as prompting an LLM to mark its response. The model can only use the information provided through retrieval, which creates several risks:

1. **Missing document markings:** Retrieved excerpts may omit banners, headers, footers, or metadata containing the source document's classification and handling controls.
2. **Separated portion markings:** Document processing may separate a portion marking from the paragraph, table, or section it governs, causing under- or over-marking.
3. I**ncomplete context:** An excerpt may exclude surrounding information needed to determine classification, dissemination controls, or the effect of combining multiple facts.
4. **Processing errors:** OCR, parsing, and chunking may remove, corrupt, or misassociate markings before the content reaches the model.
5. **Newly classified synthesis:** Combining or summarizing correctly marked sources may reveal information requiring a higher classification or additional controls.

Addressing these risks requires a systematic approach to handling markings. For a platform processing classified information, proper handling starts at ingestion, the moment classified content enters the platform. Classification banners, portion markings, dissemination controls, and handling caveats may appear in headers, footers, paragraphs, tables, images, or metadata and need to be accurately detected. The system must identify them before the document is divided into passages for indexing and retrieval. It must also account for inconsistent formatting, scanned documents, and OCR errors.

Once classification markings are detected, the platform must store the markings as permissions on the source documents. A portion marking should not be separated from its passage, table, or image during indexing. The platform must enforce these permissions before information reaches the model. An LLM should never receive content that the user requesting the answer is not authorized to access. A good platform

will implement access controls that ensure that users and AI workflows can only retrieve information they are authorized to see.

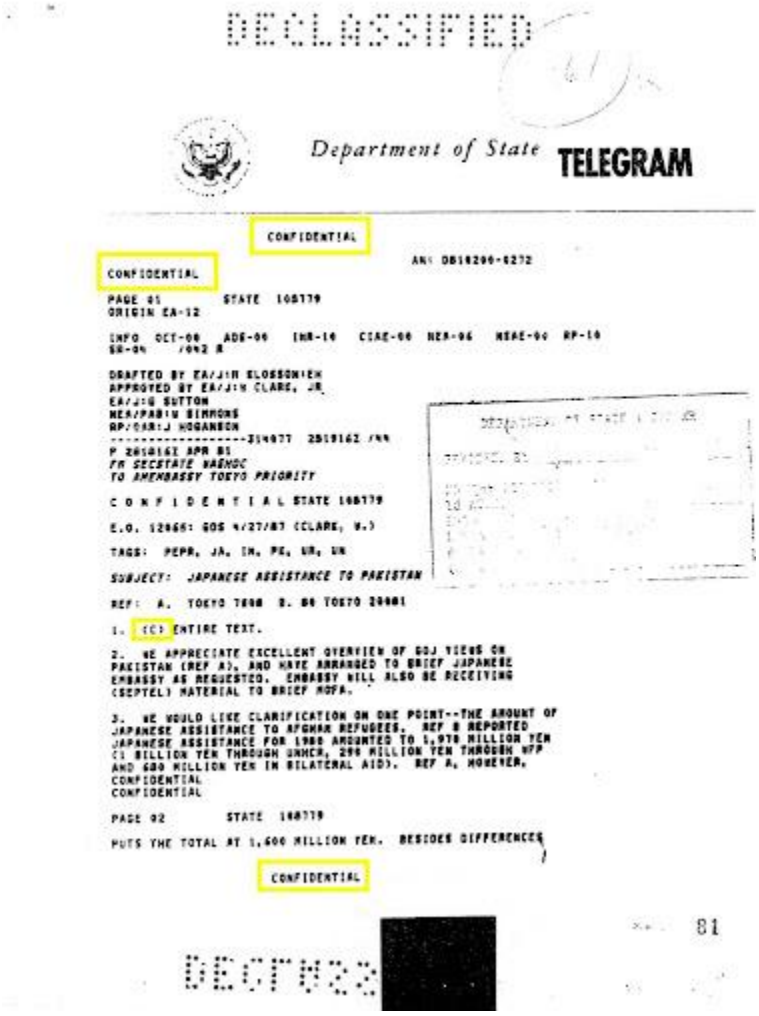
Retrieval and generation must carry these markings forward. For the Intelligence Community, ICD 505 states this as a requirement: AI systems must carry forward the handling requirements of the data they analyze, an obligation that sits with the system rather than the analyst. When the platform selects a source to use for an answer, it should retrieve the source content together with its document and portion markings. By doing so, the platform is able to maintain the connection between generated content, citations, markings and the source content. This enables the system to then properly propose markings on its generated content.

Finally, the platform needs to incorporate human review as a core tenet. Reviewers should see the proposed marking, the passages behind it, and their original markings. They must be able to confirm or correct the recommendation without reconstructing the entire retrieval process manually. This is exceptionally important for cases where algorithms may be unsure about specific proposed markings due to OCR uncertainty, confusing document layouts, or other degenerate situations.

Reliable Marking Starts with Detection in the Source Material

If the system misclassifies existing document markings, every downstream step suffers.

As previously mentioned, the first step in reliable classification handling begins with correctly identifying the markings already present in source material. If those markings are missed or misread, every downstream step inherits the error, from passage resolution to the portion marking collation. The Legion Intelligence Platform detects and extracts portion and document markings by pairing model-based extraction with deterministic algorithms for recognizable marking syntax. The model can interpret context that fixed rules may miss, while deterministic checks can catch likely omissions.



DECLASSIFIED

Department of State **TELEGRAM**

CONFIDENTIAL

CONFIDENTIAL

PAGE 01 STATE 168779
ORIGIN EA-12

INFO OCT-00 ADS-00 INR-10 CIAE-00 NSA-06 NSAE-00 RP-10
SS-04 /042 R

DRAFTED BY EA/JA:N GLOSSNICK
APPROVED BY EA/JA:N CLARE, JR.
EA/JA:G SUTTON
NEA/PAB:N SIMMONS
RP/DAR:J HOGANSON

P 2818162 APR 81
FM SECSTATE WASHDC
TO AMEMBASSY TOKYO PRIORITY

C O N F I D E N T I A L STATE 168779

E.O. 12065: GDS 4/27/87 (CLARE, W.)

TAGS: PEPR, JA, IN, PK, UR, UN

SUBJECT: JAPANESE ASSISTANCE TO PAKISTAN

REF: A. TOKYO 7898 B. 80 TOKYO 20401

1. (C) ENTIRE TEXT.

2. WE APPRECIATE EXCELLENT OVERVIEW OF GOJ VIEWS ON PAKISTAN (REF A), AND HAVE ARRANGED TO BRIEF JAPANESE EMBASSY AS REQUESTED. EMBASSY WILL ALSO BE RECEIVING (SEPTEL) MATERIAL TO BRIEF MOFA.

3. WE WOULD LIKE CLARIFICATION ON ONE POINT--THE AMOUNT OF JAPANESE ASSISTANCE TO AFGHAN REFUGEES. REF B REPORTED JAPANESE ASSISTANCE FOR 1980 AMOUNTED TO 1,970 MILLION YEN (1 BILLION YEN THROUGH UNHCR, 200 MILLION YEN THROUGH WFP AND 680 MILLION YEN IN BILATERAL AID). REF A, HOWEVER,
CONFIDENTIAL
CONFIDENTIAL

PAGE 02 STATE 168779

PUTS THE TOTAL AT 1,600 MILLION YEN. BESIDES DIFFERENCES

CONFIDENTIAL

81

DECLASS [redacted]

Internally, Legion optimized and validated its detection algorithms against both sampled declassified documents and synthetically generated documents with portion markings. Legion ran its production detection logic over these documents without showing the system the correct answers. The test included passages with portion markings, document-level banners, and no markings at all. After model extraction and deterministic checks were complete, the detected markings were normalized into standard tags such as classification levels and handling caveats and compared with the human-reviewed labels. This measured both how many expected markings Legion found and how often its detections were correct.

Across 14,831 human-labeled passages, Legion found nearly 96% of the expected classification and handling markings. That is recall, the ability to avoid missed markings. Of the markings Legion detected, 94% were correct. That is precision, the ability to avoid incorrect added markings. Together, these results produced a 94.7% F1 score.

The results are not perfect, and that is partly by design. The benchmark went beyond straightforward markings such as **(S)** or **(TS)** to probe the system's limits. It included long markings with multiple handling rules. One example combined **TOP SECRET** with release permissions, dissemination controls, and special-access program names requiring normalization. Other examples probed subtle distinctions, such as NATO appearing as a control versus as a release destination, and passages that merely mentioned terms such as EXDIS or displayed a marking as an example. These cases tested whether Legion could separate the markings applied to a passage from similar-looking language appearing within its content.

Even under these demanding test conditions, the remaining errors still matter in a final product. A missed marking can leave information below its required level of protection, while an unwarranted added marking can restrict access, delay sharing, and create unnecessary rework. This is why Legion uses AI to support, not replace, authorized reviewers: the system identifies likely markings and potential ambiguities, while the analyst remains responsible for confirming the final markings. Legion solves for these edges in a single review interface, enabling an authorized user to quickly evaluate the evidence and confirm or correct markings without leaving the workflow.

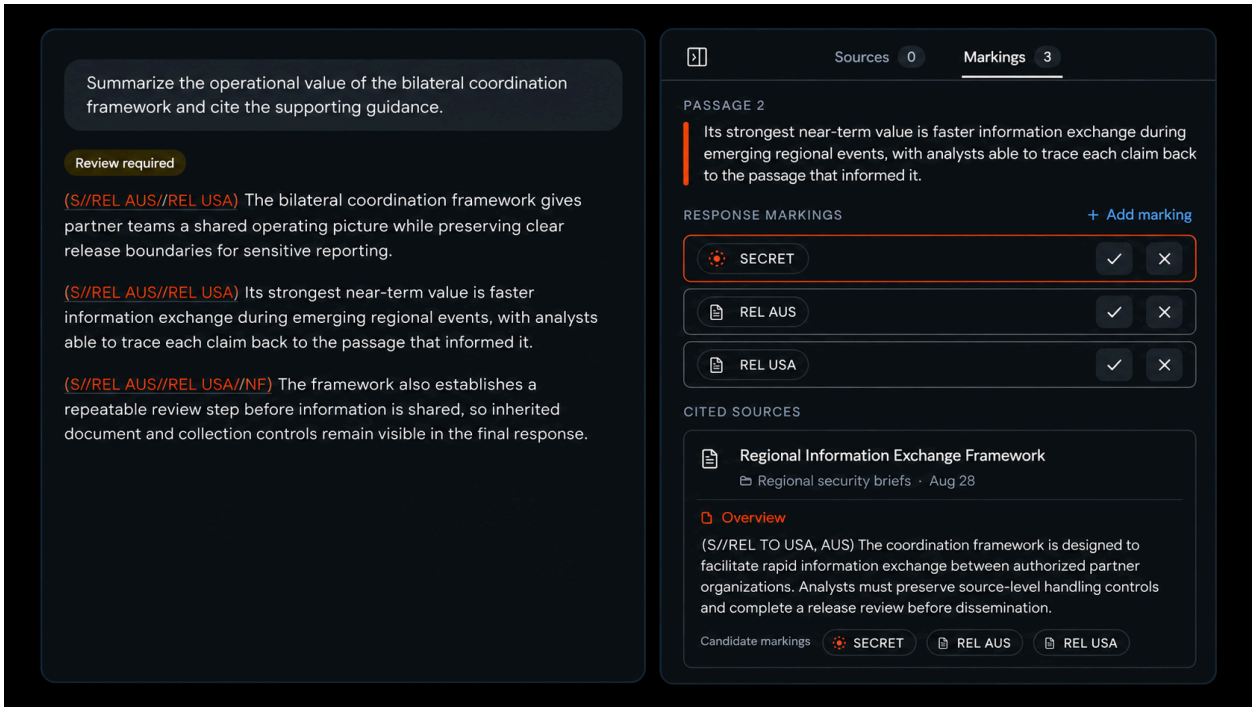
Marking Review Requires a First Class Solution

A user needs enough evidence in front of them to confirm or correct a proposed marking in seconds.

Classification marking review should be a first-class part of the experience surrounding generated content. Legion presents each proposed marking alongside the supporting evidence, allowing users to inspect the relevant source material and its existing document- and passage-level markings without reconstructing the model's reasoning from scratch.

A source marking describes how retrieved information is handled, whereas a proposed response marking applies to newly generated text. The two may differ, particularly when an answer combines information from multiple sources. Legion brings the answer, candidate markings, and supporting evidence into one streamlined review experience so an authorized user can confirm or correct the result with the relevant context in view.

By making that context immediately visible, Legion reduces the manual work and cognitive load required to reconstruct how an answer was produced. The end result is AI-generated text that arrives ready for marking review, rather than an unmarked answer that creates a new research task for the analyst.



Ultimately, effective AI in classified environments is measured not only by the quality or speed of an answer, but also by whether that answer can move safely from generation to decision. By detecting existing markings, proposing appropriate markings for generated content, and presenting the supporting evidence in a governed, traceable review workflow, Legion shortens the distance between insight and action while keeping final authority with the operators using the platform. The result is not autonomous classification, but rather a practical system that helps analysts protect information, share it appropriately, and act quickly while the answer still matters.

Three Questions Reveal Whether a System Was Designed for Classified Work

Any vendor will say they handle classification. Three questions will tell you whether a system was designed to handle classified material.

Does the system preserve portion markings?

A single document may contain portions with different classifications or dissemination controls. A system that only extracts document-level information leads to overclassification and misses the nuance of portion markings.

Can users connect LLM generated information to its supporting evidence?

A citation produced by an AI platform should not only link the relevant generated context with its source passages, but also highlight the original portion markings associated with them. Pointing to a hundred-page document doesn't help an analyst understand specific content classification.

Are markings inspectable, traceable, and correctable?

In a platform's generation, users should see what markings are proposed and why, and the system should preserve a record of the final decision.

Continue the series

Follow Legion Intelligence on LinkedIn: linkedin.com/company/legion-intel

Explore all Command Papers: legionintel.com/command-papers

Ready to talk: https://www.legionintel.com/contact