INFLUENTIAL BUT MISALIGNED: LLM EVIDENCE DRIVES PREDICTIONS YET MISSES WHAT INSTITUTIONS CLASSIFIED

Anonymous authors
Paper under double-blind review

ABSTRACT

High-reliability security document classification requires knowing not only how accurately language models predict but also what evidence supports their decisions. Prior work has evaluated the predictive influence of model rationales and their agreement with human annotations, but research annotations alone cannot establish alignment with records of actual classification practice. We use portion markings in declassified documents released through the CIA Electronic Reading Room as a comparison reference. Portion markings record classification levels for document spans; where the highest-level rule applies, the highest-classified portions are directly linked to the document label. Using texts with classification markings removed, we evaluate the predictive influence and selection consistency of self-rationales from three open-weight language models and compare them with institutional markings in the portion-recovered subset. In a stratified sample, deleting the top self-rationale changed predictions 10.5–12.4 percentage points more often than length-matched random deletion. Rationale selection was stable under identical inputs but varied with prompt wording and model choice. In the portion-recovered subset, model rationales overlapped institutional portions less than a within-document random-position baseline (lift 0.65–0.75), and we found no evidence that deleting classified portions changed predictions more often than amount-matched random deletion. This divergence also left a positional and lexical signature. Llama rationales appeared near the beginning of documents more often than institutional portions and contained classification and secrecy vocabulary 26.6 percentage points more often in a sentence comparison matched for position and length. This pattern is consistent with a preference for sentences that overtly signal classification. The predictive influence of model rationales and their alignment with institutional span-level classification records therefore require separate evaluation.

1 INTRODUCTION

In high-reliability security document classification, accuracy alone cannot make a model's decision trustworthy. Models can achieve high accuracy through shortcuts (Eshuijs et al., 2025), and models with comparable accuracy can differ in their agreement with human evaluation rationales (Wang et al., 2026). The evidence supporting predictions must therefore also be evaluated.

Brandl & Eberle (2026) evaluated extractive self-explanations through agreement with human rationales and deletion effects, and Fayyaz et al. (2024) compared them with human rationales across tasks. Yet those human rationales are annotations written or selected for research. Interpretability research distinguishes whether an explanation is faithful to model behavior (faithfulness) from whether it is plausible to people (plausibility) (Jacovi & Goldberg, 2020). We add a third axis: Is the evidence cited by a model faithful to the decision procedure that produced the label?

We use portion markings in declassified documents released through the CIA Electronic Reading Room. Portion markings record the classification level assigned to each section of a document. Under the general rule of 32 CFR §2001.21(b) for marking U.S. national-security information, a document's overall classification follows its highest-classified portion (Information Security Over-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

sight Office, National Archives and Records Administration, 2025). Where this rule applies, the highest-classified portions are directly linked to the document label; their markings are records produced in the course of work, not annotations selected afterward for this study.

Using declassified documents for evaluation requires selecting analyzable documents from a large collection with heterogeneous scan and OCR quality, and retaining classification markings, which reveal the answer, as evaluation information while removing them from the model input. We construct the data in three stages: document selection, marking removal, and portion-span alignment.

We evaluate three things on the resulting dataset. (1) *Predictive influence of self-rationales*: we sequentially delete top-ranked rationales and compare against random deletion of the same amount. (2) *Consistency and sharing of rationale selection*: we measure stability under repeated runs and prompt rephrasings, overlap across models, and cross-deletion effects. (3) *Alignment with institutional portions*: we evaluate agreement between model rationales and recovered institutional portions and the effect of deletion, and analyze selection tendencies by position, vocabulary, and formatting.

Our contributions and main observations are threefold. First, we recover portion markings as institutional records connecting span-level classifications to a document label, providing a reference distinct from annotations made for research. Second, in a stratified sample of three models, deleting the top self-rationale changed predictions 10.5–12.4 percentage points more often than equal-length random deletion. In the separate portion-recovered subset, model rationales overlapped institutional portions less than random-position spans (lift 0.65–0.75), and the excess prediction change after deleting C/S/TS portions had confidence intervals containing zero for all three models. Third, the models did not select identical sentences, and a preference for early document positions was observed. In the paired-sentence analysis for Llama, classification and control vocabulary appeared more often in model rationales than in institutional portions.

2 Related work

Faithfulness to the model. ERASER (DeYoung et al., 2020) separated agreement with human rationales from prediction changes under rationale deletion, and Madsen et al. (2024) measured the faithfulness of LLM self-explanations by removing the words the model identified as important. Validation against counterfactual inputs has extended to natural language explanations (Atanasova et al., 2023), simulatability of model outputs (Chen et al., 2024), and concept-level interventions (Matton et al., 2025); Bueno et al. (2026) compared the deletion effects of SHAP-selected sentences and LLM self-rationales across models, and FaithLM (Chuang et al., 2026) measures prediction responses to hints that contradict the explanation. Chain-of-thought explanations have also been shown to omit the factors a model actually relied on (Turpin et al., 2023). Input interventions have been criticized for measuring output-level self-consistency rather than a model's internal reasoning (Parcalabescu & Frank, 2024), and we likewise read the changes observed under intervention as output-level behavioral responses.

Plausibility to annotators. Human rationales were introduced as span annotations supporting label decisions in text classification (Zaidan et al., 2007) and later extended to explanations and highlights for natural language inference (Camburu et al., 2018) and evidence spans in clinical trial reports (Lehman et al., 2019); datasets differ in format, using highlights, free-text explanations, or structured explanations (Wiegreffe & Marasović, 2021). Fayyaz et al. (2024) compared prompt-extracted self-rationales and attention- or gradient-based rationales against human annotations, and Brandl & Eberle (2026) analyzed the alignment between extractive self-explanations and human rationales in sentiment classification, forced labor detection, and climate-claim verification. These references are annotations written or selected for research; we use portion markings recorded in the course of classification as our comparison reference.

Shortcuts and prompt sensitivity. Models can rely on cues associated with a label instead of the information central to the task (Geirhos et al., 2020), and Eshuijs et al. (2025) analyzed how an actor name in a movie review acts as a shortcut for the sentiment label. Zhao & Daumé III (2025) proposed PEX, which measures how much more a free-text explanation supports a given prediction than its opposite, and Errica et al. (2025) measured the sensitivity and consistency of classification predictions under prompt wording. Performance can shift substantially with the format or wording of a meaning-preserving prompt (Sclar et al., 2024), and a single prompt has been reported to be

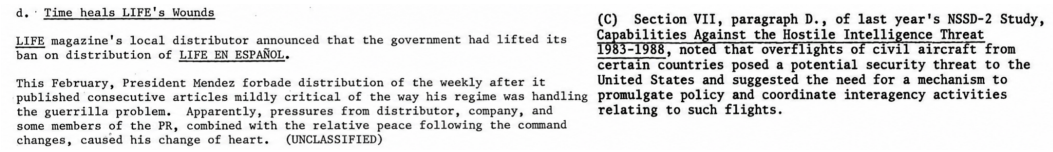
an unreliable summary of model performance (Mizrahi et al., 2024). We measure the stability of rationale selection across repeated runs, rephrasings, and models, and ask whether the selection bias is shared.

Classification of sensitive documents. Work using declassified records has automatically predicted the classification status of large diplomatic cable collections (Souza et al., 2016), semi-automated classification has been studied to support the review of sensitive documents in freedom-of-information processes (Berardi et al., 2015), and DataSIR introduced a benchmark for recognizing categories of sensitive information under diverse transformations (Mo et al., 2025).

3 DECLASSIFIED DOCUMENTS AND PORTION MARKINGS

We use documents released through the CIA Electronic Reading Room. These scanned records include intelligence reports, diplomatic documents, and interagency correspondence created in the course of actual operations.

A document-level classification does not indicate the same protection level for every span. Portion markings record classification status span by span. The marks `(C)`, `(S)`, and `(TS)` denote Confidential, Secret, and Top Secret, respectively, and `(U)` denotes Unclassified. Figure 1 shows two documents with portion markings.


d. Time heals LIFE's Wounds

LIFE magazine's local distributor announced that the government had lifted its ban on distribution of LIFE EN ESPAÑOL.

This February, President Mendez forbade distribution of the weekly after it published consecutive articles mildly critical of the way his regime was handling the guerrilla problem. Apparently, pressures from distributor, company, and some members of the PR, combined with the relative peace following the command changes, caused his change of heart. (UNCLASSIFIED)

(C) Section VII, paragraph D., of last year's NSSD-2 Study, Capabilities Against the Hostile Intelligence Threat 1983-1988, noted that overflights of civil aircraft from certain countries posed a potential security threat to the United States and suggested the need for a mechanism to promulgate policy and coordinate interagency activities relating to such flights.

Figure 1: Examples of portion-level classification markings appearing after and before the text. Left: "Biweekly Political Review" by the U.S. Embassy in Guatemala. Right: "National Policy for Overflights of Foreign Civil Aircraft" by the Interagency Group/Countermeasures.

E.O. 13526 requires the original and derivative classification procedures and the marking of portion-level classification status. Under the general overall-marking rule of 32 CFR §2001.21(b), a document's overall classification follows its highest-classified portion. Marking practice varied with the rules and agency guidance in force when a document was created; portion marking and trailing paragraph marks also appear in E.O. 12356 and a 1979 CIA guidance memorandum (President of the United States, 1982; Central Intelligence Agency, Office of Security, 1979).

Where this highest-level rule applies, let $M_i = \{(s_{ij}, c_{ij})\}_{j=1}^{n_i}$ be the complete set of portion markings in source document i. Here s_{ij} is the text span aligned with a marking, and $c_{ij} \in \{\mathrm{U}, \mathrm{C}, \mathrm{S}, \mathrm{TS}\}$ is its classification, ordered as $\mathrm{U} < \mathrm{C} < \mathrm{S} < \mathrm{TS}$. The document's overall level y_i and the set P_i of highest-level portions are related by

$$y_i = \max_{(s,c)\in M_i} c, \qquad P_i = \{s : (s, c) \in M_i,\ c = y_i\}.$$

Under this rule, retaining only P_i preserves the highest marked level, while the remaining portions cannot preserve that level once all of P_i is removed. The compilation exception in 32 CFR §2001.24(g) and differences among historical marking rules prevent us from applying this equation retrospectively to the entire collection.

The evaluation collection contains 37,862 documents, and from 2,299 of them (6.1%) we recovered 7,773 portions that pass the quality criteria. Classification distributions of documents and portions, and the characteristics of the subset, are given in Appendix A.

4 PIPELINE

The pipeline selects documents suitable for analysis, removes classification markings from model inputs, and aligns separately preserved source markings with text spans (Figure 2).

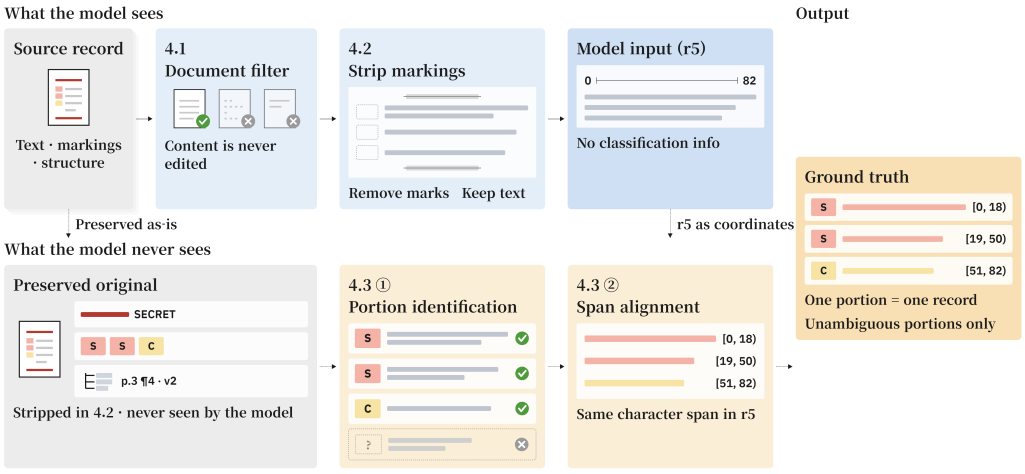


Figure 2: Construction of model inputs and institutional references. Document selection and marking removal produce the baseline input for all experiments. Portion-span alignment additionally uses the source preserved separately from this text.

Document selection. Comparing records with different scanning and OCR quality without screening them would introduce recognition errors and incorrect reading order into the model inputs (Zhang et al., 2025; Ouyang et al., 2025). We first excluded records that were too short or too long, then assessed lexical quality by the proportion of words absent from a dictionary. All-uppercase cable reports followed a separate lexical criterion. We also used word coordinates to exclude documents at high risk of mixed reading order, including multi-column layouts. These checks selected documents without rewriting the preserved source text or coordinates.

Marking removal. We removed banners and portion markings that reveal a document's classification, as well as warnings and distribution controls that disclose handling status. Because the same classification words may occur in ordinary text, removal considered page position, line form, and surrounding phrases rather than the word alone. For example, `SECRET` in `US SECRET SERVICE` is part of an agency name and is retained. When a marking shared a line with body text, we removed only the marking span rather than the entire line. We recorded removed spans and the rules applied, then checked for residual markings and over-deletion through an independent scan and revision comparison. The final r5 inputs contain no character changes beyond deletions.

Portion-span alignment. The location of a portion marking need not be the text span to which its classification applies. We located markings with source page, paragraph, line, and word coordinates, then set boundaries within each paragraph according to whether a marking preceded or followed the text. Candidates with ambiguous reading order or attachment direction were excluded. Accepted source portions were mapped to the marking-free text through exactly matching character spans, excluding deleted characters from the correspondence. Their labels and mapped spans were preserved separately from model inputs as the institutional reference. Detailed selection and removal criteria are given in Appendix B, and portion construction and validation in Appendix C.

5 Experiments

We measure the change in predictions caused by deleting rationales, examine whether rationale selection persists across conditions and models, and then compare it with institutional portions.

Common setup. We use three open-weight models of comparable scale: Llama 3.3 70B (Meta AI, 2024), Qwen 2.5 72B (Qwen Team, 2024), and Llama 3.1-based Nemotron 70B (NVIDIA, 2024). Each model receives the full r5 text constructed in Section 4 and predicts the document's original classification as one of Unclassified, Confidential, Secret, or Top Secret. Neither the document label nor the labels and locations of institutional portions are included in the input. Generation uses temperature 0.1 and top-p 0.5. We report classification performance under both the original

four-class and the grouped three-class scheme that merges Top Secret into Secret. Self-rationale and institutional-portion deletion effects use the grouped three classes; label agreement and prediction changes under repetition and cross-deletion retain the original four classes.

Along with its predicted classification, each model quotes the three sentences most decisive for its decision, ordered by importance. We define these as self-rationales and require short verbatim excerpts from the input text rather than newly written explanatory sentences. Returned quotes are aligned to character spans against the source text, and the outcome and method of each alignment are recorded. Rationale overlap and position are measured on aligned spans, and deletion experiments remove the sentence span containing each quote. Because institutional portions use the same r5 coordinates, the overlap between the two kinds of evidence can be compared directly. All brackets give 95% confidence intervals from 10,000 document-level paired bootstrap resamples. Common experimental settings are given in Appendix D, and the detailed methods and results of each experiment in Appendix E.

Classification performance on marking-free documents. Table 1 reports exact four-class and grouped three-class accuracy and macro-F1 on all 37,862 r5 documents after classification markings were removed and before any additional self-rationale or institutional-portion deletion. Under both scoring schemes, model accuracy is below the always-Secret baseline, whereas macro-F1 is higher. The analyses below distinguish whether a model predicts the label correctly from whether its cited rationales influence its prediction.

Table 1: Four- and three-class classification performance on marking-free documents (37,862 documents, before rationale or portion deletion). Three-class scoring merges Top Secret with Secret. The always-Secret baseline is shown under both schemes; nine empty Qwen responses are scored as errors.

	Four-class		Three-class	
Model	Accuracy	Macro-F1	Accuracy	Macro-F1
Llama	0.4625	0.3284	0.6306	0.4214
Qwen	0.4626	0.3205	0.6311	0.4285
Nemotron	0.4250	0.3170	0.5823	0.3942
Always Secret	0.5127	0.1695	0.7505	0.2858

5.1 Do self-rationales influence predictions?

This section evaluates whether the self-rationales a model provides are behaviorally linked to its predictions. We sequentially delete the top rationales selected by Llama and compare against random sentence deletion matched for deleted length within the same document. We define the difference in prediction-change rates between the two conditions as the *excess prediction-change rate*.

Table 2: Magnitude of Llama self-rationale deletion and position controls (three-class scoring). Excess denotes the rate difference between rationale deletion and random deletion. Full 95% confidence intervals and condition-specific denominators are reported in Appendix Table 14.

Deleted rationales	Rationale deletion change rate (%)	Random deletion change rate (%)	Excess change rate (%p)	Position-matched excess (%p)
1	16.43	5.65	10.78	9.18
2	20.48	7.57	12.91	11.00
3	23.15	9.01	14.14	12.56

Self-rationale deletion induces consistently larger prediction changes than random deletion. As Table 2 shows, deleting only the first self-rationale changes predictions 10.78%p more often than deleting a length-matched random sentence. The point estimate of the excess change rate increases with the number of rationales deleted, and its confidence interval excludes zero at every

deletion level. Even among documents whose predicted label is preserved, the newly selected first rationale is replaced by a sentence outside the previous top rationales far more often than under random deletion. A stable label therefore does not imply a stable rationale. Because long-context models can behave differently depending on where relevant information sits (Liu et al., 2024), we also added a control that deletes non-rationale sentences of similar length and position. Position matching reduces the excess change rate somewhat, but it remains positive at every deletion level.

The same trend is reproduced across the three models. As Figure 3 shows, the excess prediction-change point estimates range from 10.5–12.4%p after deleting the first self-rationale to 14.3–16.1%p after deleting all three. At every deletion level for each model, self-rationale deletion changes predictions more often than length-matched random deletion, and the 95% confidence interval for each excess rate excludes zero. The point estimates increase with deletion scope, showing that this behavioral response is not limited to Llama.

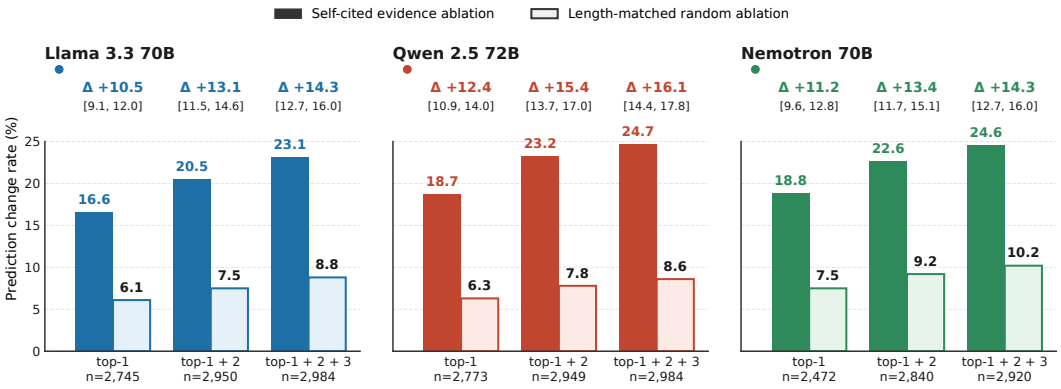


Figure 3: Cumulative self-rationale deletion effects per model on the stratified sample (three-class scoring). Filled bars show prediction-change rates (%) for self-rationale deletion and hatched bars for length-matched random deletion.

5.2 How consistent and shared is rationale selection?

The deletion effect of self-rationales appears in all three models, but whether the sentences supporting those decisions are the same across conditions and models is a separate question. We check the consistency of selection under identical-prompt repetition, meaning-preserving rephrasing, and across models, and then measure the prediction response to deleting rationales selected by another model.

Rationale selection is stable under repeated runs but changes with wording and model. As Table 3 shows, repeating an identical prompt preserves rationale spans and sentence sets at a high level. Rephrasing the prompt while preserving its meaning lowers both agreement measures together, and the three models rarely select the same rationale set. Rationale selection is relatively stable against generation variation, but it is not invariant to prompt wording or to the model.

The cross-deletion effect depends on how much rationale the models share. Each model responds most strongly to deletion of its own rationales, but deleting rationales selected by another model still changes predictions more often than random deletion. As Table 3 shows, the excess change rate is positive for every model pair when cross-deleted spans overlap at least half of the evaluated model's own rationales, whereas its direction is not consistent when there is no overlap. Cross-deletion results alone therefore cannot establish the existence of model-independent decisive sentences.

Table 3: Rationale-selection stability and overlap-stratified cross-deletion effects. (a) Repeat denotes identical-prompt repetition; Rephrase denotes meaning-preserving prompt rephrasing; Cross-model compares the three models. (b) Excess change rates are stratified by overlap with the evaluated model's own rationales.

(a) Selection stability

Condition	Span IoU	Exact (%)
Repeat	0.893–0.944	81–92
Rephrase	0.768–0.789	63–67
Cross-model	—	1.1

(b) Cross-deletion by overlap

Overlap	Excess (%p)
None	−12.2–11.7
Partial	1.4–17.4
≥ 0.5	10.7–20.0

5.3 Do rationales align with institutional portion-level judgments?

How well do the rationales that models repeatedly select, and that influence predictions, align with the portions an institution designated for protection? On the 2,299 documents from which portion labels and text spans were extracted, we compare model quotes against institutional portions. Because the number and length of portions vary by document, we use lift: the observed overlap divided by the expected overlap when quote spans of the same length are placed at random positions within the same document. A value below 1 means less overlap with portions than random position selection.

Alignment between model rationales and institutional portions is below the random-position baseline. Overall portion lift is 0.652–0.748 for the three models, all below the random-position baseline of 1. The rationales a model selects and the spans an institution designated for protection reflect different selection criteria.

Llama quotes more often contain language that overtly signals classification. As Figure 4 shows, Llama rationales concentrate nearer the front of documents than institutional portions. Their share in the first 10% of a document is roughly twice that of institutional portions. As Table 4 shows, even when Llama quotes outside all extracted portions are paired with C/S/TS portion sentences from the same documents and matched for position and length, classification and secrecy vocabulary appears 26.64%p more often in the quotes. Control and warning phrases, reporting and transmission vocabulary, list and numbering formats, and parenthetical expressions also occur more often in Llama quotes. Before marking removal was complete, 8.2% of model rationales were residual warning phrases; these cases became candidates for refining the removal rules in Section 4.

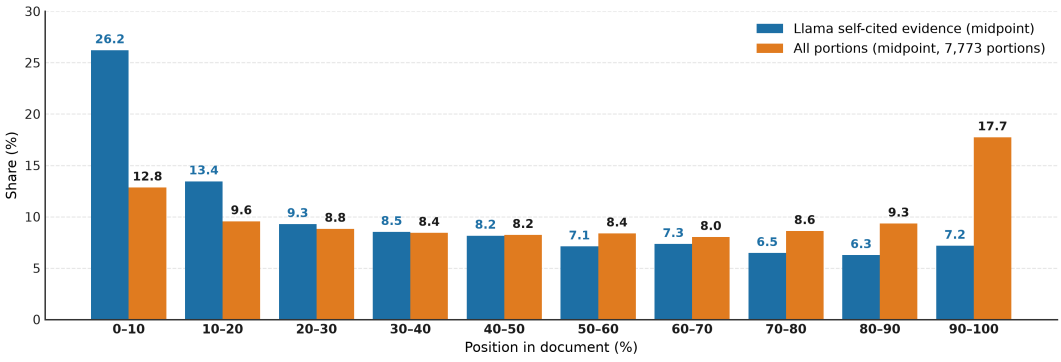


Figure 4: Position distributions of 6,493 Llama rationales and 7,773 institutional portions in 2,299 portion-bearing documents.

Institutional portions are marked whether or not the sentence reveals its classification. As Figure 4 shows, institutional portions are distributed across documents rather than concentrated at the beginning, with 17.7% falling in the last 10%. In the position- and length-matched comparison of Table 4, portion sentences also contain classification and secrecy vocabulary and control and warning phrases less often than Llama quotes. By contrast, no significant difference was detected

for semantic categories of classified information, source attribution, or uncertainty expressions. As Section 3 describes, institutional markings attach to facts that derivative classification guidance designates for protection, so a portion can be marked even when its sentence carries no cue that reveals its classification. The difference between the two kinds of evidence is especially apparent in how classification signals appear in the sentence.

Table 4: Feature occurrence differences between position- and length-matched model quotes and C/S/TS portion sentences. Positive values indicate that the feature is observed more often in model quotes. Full 95% confidence intervals are reported in Appendix Table 25.

Feature	Difference (%p)	BH q
Classification and secrecy vocabulary	26.64	<0.001
Control and warning phrases	14.11	<0.001
Reporting and transmission vocabulary	7.98	0.012
List and numbering format	6.29	<0.001
Parenthetical expressions	5.73	0.012

The predictive influence of self-rationales did not translate into alignment with institutional portions. As Table 5 shows, overall portion lift is below the random-position baseline of 1 for all three models. Matching random spans to the position distribution of model rationales, or moving them only within the same decile, still yields lift of 0.67–0.79, below 1. Relative lift for C/S/TS portions tended to be higher than for U portions, but overlap with all portions remained below the baseline. In the deletion results of Table 5, the confidence intervals of the excess change rate after C/S/TS portion deletion include zero for all three models, whereas the excess change rates after first self-rationale deletion are positive. These estimates come from separate deletion targets and control conditions. U portion deletion also showed no significant difference after multiple-comparison correction.

Table 5: Alignment with institutional portions and predictive effects of portion and self-rationale deletion. Deletion effects are three-class prediction-change rate differences from their respective length-matched random-deletion controls.

Model	Overall portion lift	C/S/TS-to-U ratio	C/S/TS portion deletion excess change (%p)	First self-rationale deletion excess change (%p)
Llama	0.652 [0.611, 0.694]	1.213 [1.048, 1.414]	−0.21 [−1.48, 1.06]	9.56 [7.92, 11.25]
Qwen	0.748 [0.708, 0.788]	1.117 [0.986, 1.276]	−0.37 [−1.43, 0.69]	10.59 [8.97, 12.22]
Nemotron	0.689 [0.647, 0.730]	1.200 [1.041, 1.393]	−0.48 [−1.85, 0.90]	12.09 [10.33, 13.84]

6 Discussion and limitations

Self-rationales can reveal which information a model responds to behaviorally, but using them as proxies for institutional judgment in security review requires separate validation against operational references.

The scope of these findings is limited to publicly released declassified documents and to the three models and prompt conditions evaluated. Institutional portions are used not as human rationale ground truth but as an external operational reference recorded in the documents, and the alignment analysis targets the set of documents for which that reference can be recovered. Deletion effects are behavioral responses to input changes and do not directly recover internal reasoning (Parcalabescu & Frank, 2024). Because we measure label changes returned by a generation API, responses in which the label is preserved while internal probabilities shift go unobserved (Siegel et al., 2024). Failure

to detect excess prediction changes after institutional-portion deletion does not show that models never use information in those portions. We also cannot verify whether the released declassified material appeared in the evaluated models' pretraining data; overlap between public benchmarks and language-model training corpora requires separate contamination measurement (Sainz et al., 2023).

7 Conclusion

This work evaluates model self-rationales against official portion records along three axes: behavioral influence, selection stability, and institutional alignment. Across three models, self-rationales affected predictions and showed some shared selection tendencies, but aligned less with institutional portions than a within-document random-position baseline. This difference also appeared in position and wording. Llama rationales occurred more often near the beginning of documents and contained classification and control language more often. Using self-rationales as evidence for institutional classification judgments therefore requires a separate check of their alignment with portion records.

AI Use Statement

In this work, the authors used ChatGPT to translate into English parts of a manuscript they had written in Korean. Generative AI tools were also used to improve the grammar, clarity, and readability of the manuscript and to render and visually refine selected figures based on layouts designed by the authors. The layouts and scientific content of those figures, including all data and numerical results, were determined and verified by the authors. All AI-assisted text and figures were reviewed by the authors to confirm that the technical content, terminology, and claims were preserved accurately and that no new scientific content was introduced. The authors take full responsibility for the final content of this paper.

Reproducibility Statement

Data sources and the institutional reference are described in Section 3, and the construction of the evaluation collection in Section 4 and Appendices B and C. Common prompts, generation conditions, and quote-alignment methods are given in Appendix D, and the samples, control conditions, metrics, and confidence-interval computations of each experiment in Appendix E. Stage artifacts together with version identification and validation records are summarized in Appendix C.2, so that processing can be traced from the source documents to the final model inputs and the institutional reference.

Ethics Statement

This work analyzes model rationales over officially released declassified records. Because public records can still contain personal data and historically sensitive information, we limit the presentation of source excerpts to what the analysis requires. Institutional portion records are used as a reference for the protection and handling decisions made at the time, and model predictions are not treated as authoritative judgments about the classification, declassification, or release of actual information.

References

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 283–294. Association for Computational Linguistics, 2023.

Giacomo Berardi, Andrea Esuli, Craig Macdonald, Iadh Ounis, and Fabrizio Sebastiani. Semi-automated text classification for sensitivity identification. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, pp. 1711–1714, 2015. doi: 10.1145/2806416.2806597.

Stephanie Brandl and Oliver Eberle. A systematic comparison between extractive self-explanations and human rationales in text classification. In *Proceedings of the 6th Workshop on Trustworthy NLP (TrustNLP 2026)*, pp. 563–583. Association for Computational Linguistics, 2026.

Ivo Bueno, Babette Bühler, Philipp Stark, Tim Fütterer, Ulrich Trautwein, Dorottya Demszky, Heather Hill, and Enkelejda Kasneci. From scoring to explanations: Evaluating SHAP and LLM rationales for rubric-based teaching quality assessment. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 7590–7606. Association for Computational Linguistics, 2026.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-SNLI: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems 31*, 2018.

Central Intelligence Agency, Office of Security. Portion Marking of Classified Information by CIA Contractors (Request for a Change), 1979. URL https://archive.org/details/cia-readingroom-document-cia-rdp94b01041r000300080001-8. November 6, 1979. CIA Reading Room document CIA-RDP94B01041R000300080001-8.

Yanda Chen, Ruiqi Zhong, Narutatsu Ri, Chen Zhao, He He, Jacob Steinhardt, Zhou Yu, and Kathleen McKeown. Do models explain themselves? counterfactual simulatability of natural language explanations. In *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235, pp. 7880–7904, 2024.

Yu-Neng Chuang, Guanchu Wang, Chia-Yuan Chang, Ruixiang Tang, Shaochen Zhong, Fan Yang, Andrew Wen, Mengnan Du, Xuanting Cai, Vladimir Braverman, and Xia Hu. FaithLM: Towards faithful explanations for large language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3802–3824. Association for Computational Linguistics, 2026.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458. Association for Computational Linguistics, 2020.

Federico Errica, Davide Sanvito, Giuseppe Siracusano, and Roberto Bifulco. What did I do wrong? quantifying LLMs' sensitivity and consistency to prompt engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1543–1558. Association for Computational Linguistics, 2025.

Leon Eshuijs, Shihan Wang, and Antske Fokkens. Short-circuiting shortcuts: Mechanistic investigation of shortcuts in text classification. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pp. 105–125. Association for Computational Linguistics, 2025.

Mohsen Fayyaz, Fan Yin, Jiao Sun, and Nanyun Peng. Evaluating human alignment and model faithfulness of LLM rationale. *arXiv preprint arXiv:2407.00219*, 2024. URL https://arxiv.org/abs/2407.00219.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Information Security Oversight Office, National Archives and Records Administration. 32 CFR 2001.21: Original classification, 2025. URL https://www.govinfo.gov/content/pkg/CFR-2025-title32-vol6/pdf/CFR-2025-title32-vol6-sec2001-21.pdf. Code of Federal Regulations, Title 32, Volume 6, July 1, 2025 edition.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205. Association for Computational Linguistics, 2020.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. Inferring which medical treatments work from reports of clinical trials. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3705–3717. Association for Computational Linguistics, 2019.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Andreas Madsen, Sarath Chandar, and Siva Reddy. Are self-explanations from large language models faithful? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 295–337. Association for Computational Linguistics, 2024.

Katie Matton, Robert Osazuwa Ness, John Guttag, and Emre Kıcıman. Walk the talk? measuring the faithfulness of large language model explanations. In *Proceedings of the 13th International Conference on Learning Representations*, 2025.

Meta AI. Llama 3.3 70b instruct model card, 2024. URL `https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct`. Accessed September 17, 2026.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024.

Fan Mo, Bo Liu, Yuan Fan, Kun Qin, Yizhou Zhao, Jinhe Zhou, Jia Sun, Jinfei Liu, and Kui Ren. DataSIR: A benchmark dataset for sensitive information recognition. In *Advances in Neural Information Processing Systems 38, Datasets and Benchmarks Track*, pp. 44298–44317, 2025.

NVIDIA. Llama-3.1-Nemotron-70B-Instruct model card, 2024. URL `https://huggingface.co/nvidia/Llama-3.1-Nemotron-70B-Instruct`. Accessed September 17, 2026.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, Jin Shi, Fan Wu, Pei Chu, Minghao Liu, Zhenxiang Li, Chao Xu, Bo Zhang, Botian Shi, Zhongying Tu, and Conghui He. OmniDocBench: Benchmarking diverse PDF document parsing with comprehensive annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24838–24848, 2025.

Letitia Parcalabescu and Anette Frank. On measuring faithfulness or self-consistency of natural language explanations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6048–6089. Association for Computational Linguistics, 2024.

President of the United States. Executive order 12356: National security information, 1982. URL `https://www.archives.gov/federal-register/codification/executive-order/12356.html`. April 2, 1982. National Archives and Records Administration.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL `https://arxiv.org/abs/2412.15115`.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10776–10787. Association for Computational Linguistics, 2023.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *Proceedings of the 12th International Conference on Learning Representations*, 2024.

Noah Siegel, Oana-Maria Camburu, Nicolas Heess, and Maria Perez-Ortiz. The probabilities also matter: A more faithful metric for faithfulness of free-text explanations in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 530–546. Association for Computational Linguistics, 2024.

Renato Rocha Souza, Flavio Codecço Coelho, Rohan Shah, and Matthew Connelly. Using artificial intelligence to identify state secrets. *arXiv preprint arXiv:1611.00356*, 2016. URL `https://arxiv.org/abs/1611.00356`.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, pp. 74952–74965, 2023. doi: 10.52202/075280-3275.

Binghai Wang, Yantao Liu, Yuxuan Liu, Tianyi Tang, Shenzhi Wang, Chang Gao, Chujie Zheng, Yichang Zhang, Le Yu, Shixuan Liu, Tao Gui, Qi Zhang, Xuanjing Huang, Bowen Yu, Fei Huang, and Junyang Lin. Outcome accuracy is not enough: Aligning the reasoning process of reward models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 41482–41508. Association for Computational Linguistics, 2026.

Sarah Wiegreffe and Ana Marasović. Teach me to explain: A review of datasets for explainable natural language processing. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1 (NeurIPS Datasets and Benchmarks 2021)*, 2021.

Omar Zaidan, Jason Eisner, and Christine Piatko. Using "annotator rationales" to improve machine learning for text categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pp. 260–267. Association for Computational Linguistics, 2007.

Junyuan Zhang, Qintong Zhang, Bin Wang, Linke Ouyang, Zichen Wen, Ying Li, Ka-Ho Chow, Conghui He, and Wentao Zhang. OCR hinders RAG: Evaluating the cascading impact of OCR on retrieval-augmented generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17443–17453, 2025.

Lingjun Zhao and Hal Daumé III. A necessary step toward faithfulness: Measuring and improving consistency in free-text explanations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15799–15813. Association for Computational Linguistics, 2025.

A CHARACTERISTICS OF THE PORTION SUBSET

Table 6 reports document-level and recovered portion-level classifications separately. Documents and portions are different counting units: a Top Secret document, for example, may contain lower-classification portions.

Table 6: Composition of the r5 evaluation collection and institutional portion references. The first two count columns summarize document-level classifications; the last summarizes the classifications of recovered spans.

Classification	All documents: count (%)	Portion subset: documents (%)	Recovered portions: spans (%)
Unclassified (U)	536 (1.4)	85 (3.7)	2,755 (35.4)
Confidential (C)	8,912 (23.5)	393 (17.1)	1,530 (19.7)
Secret (S)	19,412 (51.3)	1,104 (48.0)	3,375 (43.4)
Top Secret (TS)	9,002 (23.8)	717 (31.2)	113 (1.5)
Total	37,862 (100.0)	2,299 (100.0)	7,773 (100.0)

The portion subset used for institutional alignment consists of the 2,299 of the 37,862 r5 documents (6.1%) with recovered portion references. We compare this subset with the remaining 35,563 documents to characterize the documents in which portions can be observed. The classifications below are document-level labels, distinct from the distribution of individual portion labels in Table 6.

A.1 DOCUMENT LENGTH AND CLASSIFICATION DISTRIBUTION

Table 7: Length and classification distributions for the full collection, the portion subset, and the remaining documents. Text length is the character count of the marking-free model input; Q1 and Q3 are the first and third quartiles. Classification percentages use each column's document count as the denominator and are adjusted by the largest-remainder method to sum to 100.0%: each value is rounded down to 0.1% and the remaining mass is assigned in steps of 0.1 percentage points in order of the largest fractional part.

Characteristic	All documents	Portion subset	Remaining documents
Documents	37,862	2,299	35,563
Median length [Q1, Q3], chars	5,053 [3,944, 6,430]	4,985 [3,860, 6,365]	5,057 [3,949, 6,434]
Unclassified	1.4%	3.7%	1.3%
Confidential	23.5%	17.1%	23.9%
Secret	51.3%	48.0%	51.5%
Top Secret	23.8%	31.2%	23.3%

The portion subset is close to the remaining documents in text length but contains a higher share of Top Secret and a lower share of Confidential documents. Obtaining documents of similar length therefore does not imply a matching classification composition.

A.2 SOURCE SERIES AND TOPIC DISTRIBUTION

Source series are distinguished by the identifier format of the public archive. After removing the common prefix `cia-readingroom-document-`, identifiers beginning with `cia` are classified as CREST (CIA-RDP), `loc` as LOC, and `nsa` as NSA; the rest are grouped as numbered Reading Room identifiers, which include both purely numeric and mixed alphanumeric forms. Topic distributions are aggregated by joining existing topic tags on document identifiers.

Table 8: Distribution of source series and existing topic tags. Percentages use each collection's total document count as the denominator; source series and topics are aggregated separately. The row for the single NSA document is omitted. Totals may differ from 100% because of rounding to one decimal place and the omitted row.

Group	Item	All documents	Portion subset	Remaining
Series	CREST (CIA-RDP)	88.5%	90.6%	88.4%
Series	Reading Room numbered	9.9%	8.9%	9.9%
Series	LOC	1.6%	0.5%	1.7%
Topic	Intelligence and reconnaissance	31.5%	21.4%	32.2%
Topic	Military and weapons	24.3%	37.5%	23.4%
Topic	Politics and diplomacy	23.9%	22.3%	24.0%
Topic	Economy and industry	14.0%	11.4%	14.2%
Topic	Science and engineering	3.2%	3.0%	3.2%
Topic	Society and people	2.2%	3.3%	2.1%
Topic	Untagged	0.9%	1.1%	0.9%

The portion subset contains a higher share of military and weapons documents and a lower share of intelligence and reconnaissance documents than the rest of the collection. Institutional alignment results are interpreted as observations on a subset with this composition. Within the subset, the union of recovered portion spans covers a median of 8.9% and a mean of 16.1% of each document's text. This proportion divides the length of the aligned spans on the marking-free text by the length of that text and does not represent the coverage of all portions present in the source.

B Document quality selection and marking removal

B.1 Detailed criteria for document quality selection

Documents meeting the 2,000–10,000 character length condition undergo a layout check followed by one of two lexical decision paths. Hyphenated line breaks are joined and newlines normalized for quality computation only, reducing miscounts of validly split words; the stored text and source coordinates are unchanged.

B.1.1 Lexical metrics and branching

Let T_d denote the alphabetic token sequences of length at least 3 in the decision representation, and L_d its subsequence of tokens beginning with a lowercase letter. Let $k(w) \in \{0, 1\}$ indicate whether the lowercased word is accepted by the dictionary and its inflection and compounding rules, and let $c_d(w)$ be its case-insensitive count within the document. The document-selection metric q_{lc} and the uppercase-path metric are

$$q_{\mathrm{lc}}(d) = \frac{\sum_{w \in L_d} \mathbf{1}[k(w) = 0]}{|L_d|}, \qquad q_{\mathrm{sg}}(d) = \frac{\sum_{w \in T_d} \mathbf{1}[k(w) = 0 \wedge c_d(w) = 1]}{|T_d|}.$$

The proportion of uppercase-initial tokens is $u_d = 1 - |L_d|/|T_d|$, and documents without valid tokens are excluded. Documents with insufficient lowercase statistics are judged by the proportion of singleton out-of-dictionary words and symbol density, reducing the influence of repeated proper names.

Table 9: Common layout check and lexical criteria by document type.

Check	Applies to	Criterion
Common layout check	All documents	Excluded if the document's narrow-line ratio is at least 0.80 and at least one page is multi-column
General lexical check	$u_d \leq 0.90$ and $\lvert L_d \rvert \geq 40$	$q_{\mathrm{lc}} \leq 0.08$
Uppercase or lowercase-sparse check	$u_d > 0.90$ or $\lvert L_d \rvert < 40$	$q_{\mathrm{sg}} \leq 0.13$ and at most 5 abnormal symbols per 1,000 characters

In implementation, the OOV, singleton-OOV, uppercase-initial, and narrow-line ratios are rounded to four decimal places and abnormal-symbol density to three before threshold comparison. The permitted character set for abnormal symbols and the dictionary and supplementary dictionary configurations are specified in the reproducibility materials.

B.1.2 Multi-column layout check

We compute the horizontal extent of every line containing at least two words; pages with fewer than 10 valid lines are excluded from the layout decision. Lines narrower than 55% of the page width are defined as narrow. When narrow lines make up at least 60% of a page's valid lines, the sorted line start positions are clustered, joining adjacent points whose gaps are at most 12% of the page width. A page is judged multi-column when at least two clusters each contain at least 15% of the narrow lines and no fewer than four lines. Per-document narrow-line ratios and multi-column page counts feed the exclusion criterion of Table 9.

B.1.3 Outputs and full-collection decisions

Of the 220,535 records assessed for length, 199,051 (90.3%) meet the 2,000–10,000 character condition, and 85,558 of these (43.0%) pass the quality checks. Combining them with documents whose classification could be obtained from coordinates and applying document-level exclusions yields the final 37,862 documents.

Table 10: Quality decisions for the 199,051 documents meeting the length condition. Passing documents on the general and the uppercase or lowercase-sparse paths total 85,558.

First decision	Documents
General lexical path: pass	83,378
Uppercase or lowercase-sparse path: pass	2,180
General OOV: excluded	83,330
Multi-column layout: excluded	24,229
Uppercase path singleton-OOV: excluded	3,940
Uppercase path symbol density: excluded	1,994
Total	199,051

Table 10 counts first decisions in the code's order of application, so each document appears in exactly one row. Per-document metrics, pass status, and exclusion reasons are preserved, allowing re-adjudication under changed thresholds without reprocessing the source text.

B.2 Detailed criteria and revisions for marking removal

B.2.1 Removal scope and units

Document-level classifications are extracted from the top and bottom banner regions of the coordinate XML before removal. Classification words are normalized to uppercase letters with OCR variants allowed, while release and declassification stamps and conditional clauses are excluded. When multiple classifications are detected, the highest is taken; cases with insufficient or conflicting evidence are excluded. Labels and source markings are preserved first, and the marking characters are then removed from the model input.

Removal targets are identified by combining the positional, orthographic, and contextual conditions of Table 11. Narrow OCR fragments are processed after composite markings, and additional detection runs after page-level reconstruction. Residual portion markings in r5 are identified by positional conditions with false-positive guards, and residual banners by short line forms together with the lexical validity of neighboring tokens. Markings merged into text are excised as spans only, and residual banner classification words in r5 as tokens only.

Table 11: Signals and removal units by marking group. The deployed rules additionally include OCR variants and negative guards.

Marking group	Main signals	Removal unit
Page banners and reclassification stamps	Top or bottom position, short standalone lines, OCR variants	Line or classification token
Portion markings	Paragraph boundary, list number, sentence end, or title end; negative guards	Marking span
Exemption, control, distribution, and warning markings	Position relative to classification words; legal-citation and formulaic anchors	Matched span
Page-boundary markings	Second-pass detection on reconstructed page text	Matched span

Identification and the scope of deletion are decided separately. When OCR merges a banner and body text onto one line, the whole line is not removed even if a classification word is detected. Control phrases following a portion marking vary widely under OCR, so candidates are found mainly by positional conditions while serial numbers such as `(S/N 64-23)` and aircraft designations such as `(U-2C)` are preserved. Warning keywords are likewise confirmed together with neighboring formulaic phrases or legal citations. These conditions separate security-related expressions in ordinary text from markings that directly state the classification or handling status of a document.

B.2.2 Revisions and audits

Residuals are checked by exhaustive regex scans and fuzzy detection that share no patterns with the masking rules, and over-deletion by comparing newly deleted spans across revisions. Among 92,539 early-input model rationales, 7,573 (8.2%) were confirmed residual warning phrases, and the quotes were used to find removal candidates. Confirmed marking types were incorporated as corpus-wide rules.

B.2.3 R5 CHANGES

From r4 to r5, the marking-removal revisions changed 1,574 documents (4.16%) and removed 29,477 characters, approximately 19 per changed document. Under each marking's detection definition, banner classification-word occurrences decreased by 55.6% and portion-marking occurrences by 86.6%; these reductions do not imply that all remaining candidates are genuine markings. The changes are recorded as removed spans and the rules applied. No subsequence-check violations were found, meaning that the revisions made no character changes other than deletions, not that over-deletion of valid text was absent.

C INSTITUTIONAL PORTION CONSTRUCTION AND VALIDATION

C.1 PORTION BOUNDARY SELECTION AND INPUT-SPAN ALIGNMENT

C.1.1 MARKING DETECTION AND BOUNDARY SELECTION

Within each page of the DjVu XML, lines are sorted by vertical and horizontal coordinates, and the words within a line are joined with spaces and matched character by character against the stored marking-inclusive text. Source paragraph identifiers and the coordinates of every line and word are retained. Portion candidates are detected from parenthesized `U`, `C`, `S`, or `TS` or from full-form classification words; abbreviated forms may carry prescribed control markings joined by slashes or hyphens. Detection is distinguished from acceptance: full-form candidates whose parenthesized content does not match a regular classification word are excluded from analysis.

A marking at the start of a line with text following it is judged before-attached, and a marking that ends a line after text is judged after-attached; before-attached markings may be preceded by numeric or alphabetic list numbers. Candidates that occupy a line alone, or whose direction cannot be determined within a line, are excluded. Markings in the same XML paragraph are ordered by source position: a before-attached marking extends to the start of the next marking or the end of the paragraph, and an after-attached marking extends back to the end of the previous marking or the start of the paragraph. A marking's scope is never extended beyond its paragraph.

Confusion between list markers and `(C)` markings is resolved by examining the three neighboring lines on the same page together with horizontal coordinates: a candidate is excluded as a list item when two or more of `(A)`, `(B)`, and `(D)` are aligned within 100 coordinate units of its horizontal position. Candidates followed directly by a semicolon, or preceded within the line by numbers or clause citations, are also excluded. When attachment directions are mixed within a multi-marking paragraph, or the paragraph's lines are not contiguous in the reconstructed reading order, no boundary is accepted.

C.1.2 INITIAL ACCEPTANCE CONDITIONS

Initial portion candidates are accepted only when the XML reconstruction exactly matches the stored marking-inclusive text, the paragraph lines are contiguous, and the attachment direction is unambiguous without mixing before- and after-attached markings. The vertical center of the marking line must fall strictly between 15% and 85% of the page height. When OCR-confidence values are available, their mean over the marking words must be at least 40; missing values alone do not exclude a candidate. This confidence is an OCR-output score, not a human-verified probability of correctness.

The source portion and its aligned counterpart must each contain at least 40 characters after whitespace is trimmed, and exactly aligned characters must cover at least half of the source span. The longest trimmed aligned span must be at least 20 characters and occur exactly once in the marking-free text. Candidates identified as list or clause citations, non-regular full-form markings, or administrative stamp content are excluded.

Initial extraction detected 18,763 candidates in 4,385 of the 37,862 documents and accepted 7,773 portions from 2,299 documents. All candidates are preserved with their acceptance status, exclusion reasons, and source locations; because one candidate may carry several exclusion reasons, per-reason counts are not mutually exclusive.

C.1.3 EXACT CHARACTER ALIGNMENT AND R5 RE-ALIGNMENT

We use the exactly matching blocks of at least four characters from `difflib.SequenceMatcher(autojunk=True)`. The intersection of each block with the source portion is transferred to input coordinates by its relative position within the block, and only adjacent output spans are merged; unmatched characters are never inferred. With source span $[a_{ij}, b_{ij})$ and the partial correspondence between matching character positions denoted π_i, the institutional reference span of Section 3 is

$$s_{ij} = \{\pi_i(t) \mid a_{ij} \le t < b_{ij},\ t \in \mathrm{dom}(\pi_i)\}.$$

Deleted characters are excluded from the correspondence, and a portion split into several spans is preserved as a span list.

The initial selection under these conditions was performed on r4 inputs. Because r5 preserves the marking-inclusive source, the selected portions' labels, source boundaries, and acceptance status are retained, and the character correspondence from the source to the r5 text is recomputed. The input-span lists of 2,300 of the 7,773 portions changed, and the same portion set over 2,299 documents is used after re-alignment. The r5 re-alignment step does not re-apply the initial acceptance conditions.

Each record stores document and portion identifiers, the source page, paragraph, line, and marking coordinates and extents, and the r5 span list, text, and hash. Comparisons with model quotes use the r5 spans.

C.2 VALIDATION AND REPRODUCIBILITY

C.2.1 CHARACTER AND COORDINATE INTEGRITY

We checked duplicate portion identifiers, agreement between the candidate manifest and the acceptance list, document existence and span validity, and agreement between reconstructed text, stored text, and hashes. On the initial r4 reference of 7,773 portions, internal integrity errors and overlaps between the aligned spans of different portions were both zero. The r5 deletion-integrity check likewise found no subsequence violations, as reported above. These checks establish character- and coordinate-level consistency, distinct from the meaning of a marking or the accuracy of a boundary.

C.2.2 ARTIFACTS AND VERSION RECORDS FOR REPRODUCTION

The stage artifacts of Table 12 are linked by document and portion identifiers, and extraction settings are recorded in processing summaries. Dataset and portion files are identified by the SHA-256 of their generation reports.

Table 12: Main executables and preserved artifacts for each stage of dataset construction.

Stage	Main executables	Preserved artifacts
Quality selection	`p2_filter_length`, `p3_filter_quality`	Per-document metrics, pass status, exclusion reasons
Initial portion selection	`build_high_confidence_portions.py`	All candidates, acceptance list, source boundaries, settings summary
Initial reference integrity audit	`audit_high_confidence_portions.py`	Character, coordinate, and hash checks on the r4 reference
r5 marking removal	`s3_build_marker_dataset.py`	r5 text, removal records, generation report
r5 span re-alignment	`s4_build_derived_cpu.py`	r5 spans, text, and hashes for the same portions; generation report

Final analyses pair the r5 text with the portion file re-aligned to it, and per-version processing records allow tracing back to source locations.

D COMMON EXPERIMENTAL SETTINGS

D.1 CLASSIFICATION AND QUOTATION PROMPTS

The deployed model identifiers are `llama3.3:70b`, `qwen2.5:72b`, and `nemotron:70b`. The system prompt instructs the model to act as a security analyst reviewing declassified government documents, to classify the original sensitivity reflected in the text, and to return only the specified JSON. The default classification prompt is:

```
Classify the document into exactly one of these labels:
Unclassified, Confidential, Secret, Top Secret. Use the most
sensitive classification evidence present in the content.
Return JSON only in this format: ```json
{"category_matched": "Secret", "evidence_quotes": ["short exact
quote from the document"], "brief_explanation": "short
reason"}```. The evidence_quotes field must contain short
exact excerpts from the document text, not paraphrases. List the
3 most decisive evidence sentences, ordered from most to least
decisive (index 0 = most decisive). Do not quote these
instructions, category names, or policy descriptions as evidence.
```

The model returns a predicted classification, a list of quotes, and a brief explanation; span comparison and deletion use `evidence_quotes`. Three sentences is the number requested in the output format and does not guarantee the number actually returned or aligned. The default prompt and rephrased variants are recorded separately.

D.2 ALIGNING QUOTES TO THE INPUT TEXT

Coordinate alignment for institutional portions and alignment for model quotes use different criteria. The portion alignment of C.1 transfers exactly matching character blocks between the preserved source and r5. Model quotes may differ in whitespace or punctuation as generated, so the staged alignment below is used. The existence of approximately aligned model quotes does not mean that the text of an institutional portion was approximately recovered.

Quote alignment attempts, in order: exact match; case-insensitive match; match after normalizing whitespace, quotation marks, and dashes; and sentence-level token overlap. Normalization applies only to the search representation, and character-level position maps return each match to a span of the original input. In the final stage, we form the set of alphanumeric tokens of length at least 2 and, when a quote has at least three tokens, select the sentence with the greatest overlap among the sentences containing at least 75% of the quote's tokens. The mapping method, including this approximate stage, and any failures are recorded separately.

The quote spans used for overlap analysis are distinguished from the sentence spans used for deletion. For deletion, mapped spans are extended to sentence boundaries and overlapping deletion spans are merged; line boundaries are used instead when a quote is at most 8 characters or the extended sentence exceeds 1,200 characters. Alignment failures are not converted into judgments that no rationale exists or that a rationale disagrees with institutional portions; they are handled by the valid-sample criteria of each analysis.

D.3 GENERATION CONDITIONS AND PREDICTION REUSE

Generation uses temperature 0.1 and top-p 0.5, and documents are provided in full without prefix-length truncation. Repetition and rephrasing experiments use a maximum generation length of 512 tokens in all conditions; generation seeds cannot be fixed on the inference path used, so backend defaults apply. Actual output variation is measured through repeated runs under identical conditions, and the seeds for document sampling, random deletion, and statistical resampling are kept separate from generation seeds.

In the transition to r5, documents whose inputs changed were re-inferred, and predictions for unchanged inputs were reused with their provenance recorded. The Llama full-collection cumulative-deletion analysis therefore includes earlier predictions generated with a 256-token limit alongside

new 512-token predictions. In a controlled comparison changing only the generation limit, label agreement was 99.49% and rationale-span IoU 0.938, with no truncation by the output limit in either condition. Comparisons across separate sessions showed 98.99% label agreement and IoU 0.9365, which we distinguish from the effect of the generation limit alone. Predictions reused in the repetition and rephrasing experiments were also generated with 512 tokens in their original runs. We therefore unify the input generation at r5 but do not describe all predictions as newly produced under a single generation limit.

E Detailed methods and results of the main experiments

E.1 Cumulative self-rationale deletion and control conditions

Intervention and metric. The top k quotes ($k = 1, 2, 3$) generated at baseline are aligned to the text, and the sentence spans containing them are merged and deleted. Each condition is constructed anew from the original text; rationales regenerated after a deletion are never used as the next deletion target. With baseline prediction $\hat{y}_i$ for document i and post-deletion predictions $\hat{y}_i^{E,k}$ and $\hat{y}_i^{R,k}$ for rationale and control deletion, the excess prediction-change rate is

$$\Delta F_k = \frac{100}{|I_k|} \sum_{i \in I_k} \left(\mathbf{1}[\hat{y}_i^{E,k} \neq \hat{y}_i] - \mathbf{1}[\hat{y}_i^{R,k} \neq \hat{y}_i] \right),$$

where I_k is the valid document set of the comparison and values are in percentage points. The main analysis merges Top Secret into Secret for both predictions and labels, and accuracy differences use the same scoring. Confidence intervals use 10,000 bootstrap resamples preserving document-level pairing, with seed 7701. Individual confidence intervals across deletion conditions do not substitute for tests of differences between cumulative effects.

The prediction-change rate is the share of documents whose label differs from the baseline, regardless of correctness, whereas the accuracy difference compares the share of correct labels between rationale and random deletion on the same valid document set. Moves between two incorrect classes therefore count toward the change rate but not toward accuracy. Directional excess rates are differences in the share of all documents changing in each direction, distinct from the downward share among changed documents.

Length-matched random deletion. Candidates are sentences in the same document that do not overlap the condition's rationale spans. Candidates within $\pm 10\%$ of each rationale sentence's length are preferred, widening to $\pm 25\%$ when none exist; in both cases the tolerance floor is 5 characters. If no candidate remains, one of the five closest-length sentences is chosen, excluding candidates that overlap already-selected sentences. Randomness is fixed by seed 7701 and the document identifier, and the selection mode, unmet slots, and actual deleted character counts are recorded. Length matching therefore does not mean that exactly the same number of characters is deleted in every document.

The denominator of the full-collection Llama analysis is the 37,862 documents with valid baseline, rationale-deletion, and random-deletion predictions. Documents whose input did not change because no rationale could be aligned are included in this full denominator, distinct from the actually-deleted denominators of the three-model analysis below.

Table 13: Cumulative deletion results on the full Llama collection. The accuracy difference subtracts random-deletion accuracy from rationale-deletion accuracy. All rows use the 37,862-document denominator.

Deleted rationales	Rationale deletion (%)	Random deletion (%)	Excess change rate (pp)	Accuracy difference (pp)
1	16.43	5.65	10.78 [10.39, 11.17]	−3.84 [−4.22, −3.47]
2	20.48	7.57	12.91 [12.47, 13.35]	−5.90 [−6.32, −5.49]
3	23.15	9.01	14.14 [13.68, 14.60]	−7.61 [−8.04, −7.18]

Table 14: Full cumulative Llama results corresponding to Table 2. Brackets give 95% confidence intervals. Prediction-change rates use all 37,862 documents. Replacement-rate denominators for rationale/random deletion are 24,368/29,142, 22,373/28,521, and 21,965/28,057 for one, two, and three deleted rationales, respectively. The corresponding valid-document counts for the position-matched control are 34,595, 37,068, and 37,476.

Deleted rationales	Rationale deletion change rate (%)	Random deletion change rate (%)	Excess change rate (pp)	Replacement after rationale deletion (%)	Replacement after random deletion (%)	Position-matched excess (pp)
1	16.43	5.65	10.78 [10.39, 11.17]	62.71 [62.10, 63.32]	7.20 [6.90, 7.50]	9.18 [8.76, 9.61]
2	20.48	7.57	12.91 [12.47, 13.35]	83.33 [82.85, 83.82]	8.80 [8.47, 9.14]	11.00 [10.54, 11.42]
3	23.15	9.01	14.14 [13.68, 14.60]	100 (by definition)	9.63 [9.28, 9.98]	12.56 [12.12, 13.03]

With four classes compared directly, the excess change rates for one, two, and three deleted rationales are 12.24 pp [11.83, 12.66], 14.20 pp [13.75, 14.65], and 15.32 pp [14.83, 15.80]. Under three-class scoring, the downward share of changed predictions is 71.72%, 76.45%, and 79.84%, the last computed over the 8,764 documents with changed predictions.

Position-matched deletion. The deleted amount is matched using non-rationale sentences within ± 1 position decile of the first rationale span. Excluding 3,129 documents without a reference rationale and 138 where the control could not be constructed, baseline, rationale deletion, length-matched deletion, and length- and position-matched deletion are compared on 34,595 documents. The median deleted-length ratio of the position-matched control is 1.0054, with 90.0% falling within $\pm 25\%$ of the rationale-deletion amount. The median number of deleted sentences per reference span is 2, and the mean position shift is 0.708 deciles. This control reduces positional differences without fixing them entirely, and because bundles of sentences are deleted, the number of discontinuities in the text may differ. On this same document set, the difference between rationale deletion and length-matched deletion is 11.37 pp [10.95, 11.79], between rationale deletion and length- and position-matched deletion 9.18 pp [8.76, 9.61], and between length- and position-matched and length-matched deletion 2.19 pp [1.89, 2.49].

Three-model stratified comparison. We sampled 2,999 documents with seed 7701 according to the four-class proportions of the full collection. Each model uses its own baseline quotes with the same deletion builder and random seeds, and Llama reuses its full-collection results on these documents. Only documents with an actual deletion and three valid predictions enter each model-condition analysis, so the valid denominators differ even though the selected document set is identical.

Table 15: Cumulative deletion results of the three models on the stratified sample. All change rates use three-class scoring and are conditional on successful rationale alignment and valid predictions.

Model	Deleted rationales	Valid documents	Change rate: rationale / random (%)	Excess change rate (pp)	95% CI
Llama	1	2,745	16.65 / 6.12	10.53	[9.07, 12.02]
Llama	2	2,950	20.54 / 7.49	13.05	[11.46, 14.58]
Llama	3	2,984	23.09 / 8.75	14.34	[12.70, 15.99]
Qwen	1	2,773	18.72 / 6.31	12.41	[10.89, 13.96]
Qwen	2	2,949	23.16 / 7.80	15.36	[13.73, 17.02]
Qwen	3	2,984	24.73 / 8.61	16.12	[14.44, 17.79]
Nemotron	1	2,472	18.77 / 7.52	11.25	[9.59, 12.82]
Nemotron	2	2,840	22.61 / 9.19	13.42	[11.73, 15.07]
Nemotron	3	2,920	24.59 / 10.24	14.35	[12.71, 16.03]

Change direction. With the three classes ordered U $<$ C $<$ S, we distinguish predictions that move below and above the baseline. The excess downward and upward rates are the rationale-

minus-random differences in the share of documents changing in each direction over the full 37,862 documents, and thus differ from the downward share whose denominator is only the changed documents.

Table 16: Directional differences in change rates between self-rationale and random deletion. Individual rounding can make the sums differ from the excess prediction-change rate by 0.01 pp.

Deleted rationales	Excess downward rate (pp)	Excess upward rate (pp)
1	8.77 [8.44, 9.09]	2.01 [1.79, 2.24]
2	11.49 [11.11, 11.86]	1.42 [1.18, 1.66]
3	13.35 [12.96, 13.75]	0.78 [0.54, 1.02]

Rationale replacement under preserved labels. We analyze documents where the sentence spans of all top three baseline quotes are aligned and the new first quote after deletion is also aligned. Post-deletion spans are mapped back to original-text coordinates using the deletion records. A new span is matched to a previous rationale when the intersection length divided by the shorter span's length is at least 0.30. Previous rationale spans removed by the intervention are excluded from the matching candidates, and a new span matching no candidate is judged a replacement; replacement therefore does not mean the two spans share no characters at all.

Among valid documents whose three-class prediction was preserved, the first rationale was replaced in 15,281 of 24,368 documents after first-rationale deletion (62.71% [62.10, 63.32]), compared with 2,098 of 29,142 after length-matched random deletion (7.20% [6.90, 7.50]). The conditions differ in which documents preserve the label and map successfully, so their denominators differ. The replacement analysis excludes 6,422 documents whose top three baseline rationales were not all aligned, with further exclusions for missing or unalignable post-deletion quotes.

In a paired comparison on the 23,274 documents where the label was preserved and mapping succeeded under both conditions, the replacement-rate difference was 54.86 pp [54.18, 55.53], which is distinct from directly subtracting the two conditional rates. After rationale deletion, the previous second rationale became the first in 25.66% of cases; after random deletion, the previous first rationale was retained in 82.68%.

Input-reconstruction checks for the first-rationale condition found no mismatches over 29,373 rationale-deletion and 31,063 random-deletion documents, and coordinate checks on spans not crossing deletion boundaries found none over 29,369 and 31,063 spans respectively; these counts are the audit scope before splitting by label preservation. Rationale replacement is a response to input deletion and is not directly differenced against the variability of identical-input repetition runs.

E.2 Consistency, sharing, and cross-model deletion of rationale selection

This section separates generation variability under identical inputs, meaning-preserving prompt changes, and rationale selection across models. Repetition of the same wording provides the baseline for generation variability, and cross-wording comparison measures differences under rephrasing while the output requirements are held fixed. Cross-deletion separates the model that selects the rationales from the model that predicts, evaluating agreement of selected spans and the prediction response to removing those spans separately.

Repetition and rephrasing. We sampled 450 documents, 56–57 from each of eight strata crossing the four classes with portion availability. This is a balanced stratified sample for per-class comparison, not a sample following the full collection's class proportions. Running P0 and three rephrasings P1–P3 twice each gives eight conditions per model and 24 in total. The label list, output format, number of rationales, and quotation instruction are fixed; only the second sentence of the D.1 prompt is varied as follows, with the wording frozen before execution.

Same-wording repetition compares the two runs of each of P0–P3, giving 1,800 document-pairs per model; cross-wording comparison covers the four repetition combinations of the six wording pairs, giving 10,800 document-pairs. Quote spans are merged per response, and IoU is the intersection character count divided by the union. Only pairs with mapped rationales on both sides enter the

Condition	Core decision instruction
P0	Use the most sensitive classification evidence present in the content.
P1	Base the document's classification on the most sensitive classification evidence present in the content.
P2	For classification, use the most sensitive evidence present in the content.
P3	Among the classification evidence present in the content, use the most sensitive.

IoU; we average pairs within documents and compute confidence intervals from 10,000 bootstrap resamples of documents within strata, never treating multiple pairs of one document as independent documents. Label agreement is exact identity of the four-class strings.

Table 17: Consistency of predictions and rationales under identical-prompt repetition and rephrasing on 450 documents. Of 3,600 responses per model, those with no mapped rationale number 22 for Llama, 22 for Qwen, and 88 for Nemotron; these are not replaced with zero-agreement scores. No output was truncated by the generation limit.

Model	Repetition label agreement (%)	Repetition rationale IoU	Cross-wording label agreement (%)	Cross-wording rationale IoU
Llama	99.8	0.944 [0.934, 0.954]	93.7	0.789 [0.767, 0.812]
Qwen	99.4	0.893 [0.879, 0.906]	94.6	0.768 [0.747, 0.789]
Nemotron	99.9	0.935 [0.924, 0.945]	91.0	0.772 [0.750, 0.794]

Sentence sharing across models. In the full-collection analysis, each quote is aligned to the source and extended to sentence boundaries. Quotes pointing to the same sentence within one response are merged, and sentences are judged identical when both start and end coordinates match. The denominator is the 36,518 of 37,862 documents where every model has at least one mapped rationale. Simultaneous three-model sharing means a non-empty intersection of the three sets, distinct from each pair sharing different sentences. Full-set agreement is identity of the unordered sentence sets and is not restricted to responses whose quotes all mapped successfully.

Table 18: Cross-model rationale sentence sharing in the full-collection analysis, over the 36,518 documents where all three models have mapped rationales; sentences with identical source start and end coordinates count as shared. Simultaneous sharing requires at least one sentence in the three-way intersection.

Comparison	At least one shared sentence (%)	95% CI
Llama–Qwen	58.8	[58.3, 59.4]
Llama–Nemotron	72.8	[72.3, 73.3]
Qwen–Nemotron	56.8	[56.3, 57.3]
All three models	40.0	[39.5, 40.5]

Confidence intervals in Table 18 use 10,000 document bootstrap resamples. The random baseline draws 5,000 valid documents and, 1,000 times, samples without replacement as many sentences per model as were actually mapped in each document; the probability of a three-way intersection is 1.22%, a factor of 32.7 below the observed rate. This baseline preserves the number of selected sentences but not their lengths or positions. Relaxing the identity criterion to span IoU ≥ 0.7 raises pairwise at-least-one-sentence sharing to 65.7% for Llama–Qwen, 78.5% for Llama–Nemotron, and 62.3% for Qwen–Nemotron. The quote-span IoU of the repetition experiments and cross-model sentence-set agreement are different metrics.

Cross-deletion construction and denominators. On the 450 documents, each model's first P0 run is frozen as its baseline, and the sentence spans containing its top three quotes are merged and deleted. Nine conditions provide the three deletion inputs to each evaluated model, plus three

Table 19: Prediction-change rates under cross-model rationale deletion. Rows are the predicting models and columns the models whose rationales are deleted; bold values are self-rationale deletion. The random column averages controls matched to the deleted amount of each rationale source.

Evaluated model	Llama	Qwen	Nemotron	Random
Llama	**25.7**	23.7	22.8	10.3
Qwen	21.9	**26.5**	20.8	9.0
Nemotron	25.8	26.5	**27.2**	12.3

conditions with the common random input. The 55 documents whose r5 input changed were re-inferred across all 12 conditions, and the 395 with unchanged inputs reuse same-condition predictions. Change rates and their differences in this analysis keep the original four classes and are not numerically compared with the three-class rates of E.1.

In Table 19, each cell of the Llama and Qwen deletion columns covers 447 documents and the Nemotron column 441. The random input comes from the top-three length-matched condition of the full-collection Llama experiment, excluding documents where it intrudes on the evaluated model's own rationales; the valid denominators of the random column are therefore 441 for Llama, 326 for Qwen, and 382 for Nemotron, and it is not a control matched to every source model's deletion amounts. Differences between two conditions always use their common valid documents. Confidence intervals in Table 20 use 10,000 document-level paired bootstrap resamples with seed 7701.

Table 20: Differences between self- or cross-rationale deletion and random deletion, computed on common valid documents.

Evaluated model	Source of deleted rationales	Common valid documents	Excess change vs. random (pp)
Llama	Llama	441	12.70 [8.39, 17.23]
Llama	Qwen	438	8.90 [4.79, 13.01]
Llama	Nemotron	433	8.31 [4.16, 12.47]
Qwen	Llama	326	13.80 [9.51, 18.40]
Qwen	Qwen	323	20.12 [15.17, 25.39]
Qwen	Nemotron	319	17.24 [12.54, 22.26]
Nemotron	Llama	382	7.07 [2.09, 12.04]
Nemotron	Qwen	379	8.44 [3.43, 13.46]
Nemotron	Nemotron	374	9.63 [4.28, 14.97]

Deletion-amount matching. When comparing self- and cross-rationale deletion, we restrict to documents where the cross-deletion character count divided by the self-deletion count lies in 0.85–1.18. The point estimates of all six comparisons are non-negative, but every confidence interval includes zero (Table 21). The observation in Table 19 that self-rationale deletion is largest does not by itself establish amount-independent model specificity.

Comparison by rationale overlap. The intersection character count between the cross-deleted spans and the evaluated model's own rationale spans is divided by the total character count of the latter. Documents are split into zero overlap, partial overlap below 0.5, and overlap at least 0.5, and within each stratum change-rate differences are computed on common documents where the random condition is also valid; the n in Table 22 is the denominator of this paired comparison, not the stratum size before considering the random condition.

Table 21: Change-rate differences on documents where self- and cross-rationale deletion amounts are similar.

Evaluated model	Cross-rationale source	Matched documents	Self − cross change rate (pp)
Llama	Qwen	144	0.00 [−5.56, 5.56]
Llama	Nemotron	180	1.67 [−2.78, 6.11]
Qwen	Llama	144	4.86 [−1.39, 11.11]
Qwen	Nemotron	125	5.60 [−1.60, 12.80]
Nemotron	Llama	180	2.78 [−2.22, 8.33]
Nemotron	Qwen	126	1.59 [−5.56, 9.52]

Table 22: Excess change rates by overlap between the evaluated model's own rationales and the cross-deleted spans. This is a descriptive comparison of different document strata, not an experiment manipulating only overlap; the mixed directions in the zero stratum are not interpreted as absence or equivalence of cross effects.

Evaluated model	Cross-rationale source	Zero overlap: excess (pp), n	Partial overlap: excess (pp), n	Overlap ≥ 0.5: excess (pp), n
Llama	Qwen	−4.4, 91	10.2, 127	13.7, 220
Llama	Nemotron	4.0, 51	3.9, 101	10.7, 281
Qwen	Llama	3.5, 57	15.9, 107	16.3, 159
Qwen	Nemotron	11.7, 68	17.4, 109	20.0, 140
Nemotron	Llama	−12.2, 41	1.4, 75	12.0, 258
Nemotron	Qwen	0.0, 86	6.7, 103	13.1, 183

E.3 Alignment with institutional portions, expression differences, and deletion effects

Overlap and position controls. Each model's top three quotes are aligned to the r5 text, and a quote counts as overlapping when it shares at least one character with a portion's text span. Portion markings themselves are absent from the model input; comparisons use the separately preserved institutional labels and r5 spans. For quote q in document d and the set $G_{d,g}$ of recovered target portion spans, let $p_{q,g}$ be the probability that a span of the same length, placed uniformly over the document's valid start positions, overlaps the set. For all portions, C/S/TS portions, and U portions we compute

$$L_g = \frac{\sum_d \sum_{q \in Q_d} \mathbf{1}[q \cap G_{d,g} \neq \varnothing]}{\sum_d \sum_{q \in Q_d} p_{q,g}}, \qquad R = \frac{L_{C/S/TS}}{L_U}.$$

L_g is the ratio of overlapping quotes to their expected count, not a character-share or a simple average of per-document lifts, and R is the ratio of classified to U lift. Spans without extracted portions are not treated as U. Confidence intervals resample observed and expected counts jointly by document, with 10,000 bootstrap resamples and seed 7701. Responses whose quotes fail to align are not converted to disagreement, so the valid documents and quotes per model are as in Table 23.

Two position controls reflecting the models' front-of-document tendency were also run. Front-matched placement selects a decile according to the observed positional distribution of rationales and places the span within it; within-document decile matching repositions the span within the decile where the rationale lies. Both preserve quote lengths. Expected overlaps in Table 24 are estimated by Monte Carlo placement and thus differ in estimation method from the analytic expectations of Table 5.

Table 23: Quotes used in the alignment computation on the 2,299-document portion subset, with per-class lifts. Confidence intervals for the overall lift and the ratio are given in Table 5.

Model	Documents with mapped quotes	Returned quotes	Mapped quotes	C/S/TS lift	U lift
Llama	2,290	6,881	6,488	0.676	0.557
Qwen	2,285	6,878	6,358	0.756	0.676
Nemotron	2,263	6,852	5,759	0.710	0.592

Table 24: Alignment analyses accounting for the positional distribution of rationales. Valid documents and quotes are 2,285 and 6,479 for Llama, 2,280 and 6,351 for Qwen, and 2,258 and 5,760 for Nemotron. Confidence intervals use 10,000 document bootstrap resamples.

Model	Position control	Overall portion lift	C/S/TS to U ratio
Llama	Front-matched	0.672 [0.631, 0.715]	1.399 [1.202, 1.636]
Llama	Within-document decile	0.728 [0.694, 0.760]	1.141 [1.021, 1.289]
Qwen	Front-matched	0.765 [0.724, 0.807]	1.276 [1.121, 1.457]
Qwen	Within-document decile	0.785 [0.755, 0.815]	1.111 [1.007, 1.231]
Nemotron	Front-matched	0.692 [0.648, 0.736]	1.348 [1.164, 1.581]
Nemotron	Within-document decile	0.736 [0.701, 0.771]	1.153 [1.027, 1.303]

Position and expression comparison. Figure 4 aggregates the midpoints of 6,493 Llama quotes and 7,773 institutional portions, divided by source length into relative positions and binned into ten intervals. The position and style analyses use a separately generated quote-alignment artifact, so quote counts differ from the alignment tallies of Table 23. Quote-count and portion-character distributions are not treated as one unit; the figure uses midpoint count shares for both groups.

The endpoint bins summarize the contrast visible in Figure 4: 26.21% of Llama quotes and 12.85% of institutional portions fall in the first decile, whereas 7.18% and 17.74%, respectively, fall in the last decile. Institutional portions here include both C/S/TS and U portions; bin shares may not sum to exactly 100% because of rounding.

The expression comparison pairs Llama quotes lying outside all extracted portions with sentences inside C/S/TS portions of the same documents. The main analysis covers 376 pairs from 326 documents with a length ratio (longer over shorter) of at most 1.25 and a position-decile difference of at most 1. Feature occurrence is computed per unit from feature dictionaries and patterns, subtracting the occurrence rate in portion sentences from that in model quotes. Confidence intervals use 10,000 document-level bootstrap resamples, and Benjamini–Hochberg (BH) correction is applied over the analyzed feature families. The five analyzed features are given in Table 25.

Table 25: Feature occurrence differences between matched sentences. Positive values mean the feature is observed more often in model quotes; they do not measure the causal influence of the feature itself.

Feature	Occurrence difference (pp)	95% CI	BH q
Classification and secrecy vocabulary	26.64	[21.98, 31.34]	<0.001
Control and warning phrases	14.11	[10.58, 17.74]	<0.001
Reporting and transmission vocabulary	7.98	[2.61, 13.39]	0.012
List and numbering format	6.29	[3.53, 9.25]	<0.001
Parenthetical expressions	5.73	[1.74, 9.87]	0.012

Differences in semantic categories of classified information, source-attribution expressions, and uncertainty expressions were not significant after multiple-comparison correction. Classification and control vocabulary here are features of the remaining text and are not the same category as the banners and portion markings defined as removal targets; the results are interpreted as selection tendencies within the remaining text.

Deleting institutional portions. We construct one condition deleting the merged text spans of C/S/TS portions and another deleting U portions. Control sentences are chosen in the same document while avoiding all institutional portions and the rationale spans of all three models, sequentially selecting sentences that reduce the gap to the target deletion amount, with ties ordered by seed 7701 per document and target; exact character-count matching is not guaranteed. Median deleted amounts are 342 characters for C/S/TS portions and 343 for their control, and 283 for both U portions and their control. The same deletion inputs are provided to all three models.

Documents with valid predictions in both the portion and control conditions number 1,893 for C/S/TS and 959 for U, with an intersection of 558, giving 2,294 documents in the full intervention set. Baselines are each model's r5 predictions from the alignment experiment. Change rates in Table 26 use three classes merging TS into S, with 10,000 bootstrap resamples and seed 7701; q applies BH correction to McNemar tests over the six comparisons crossing three models with two portion types.

Table 26: Prediction-change rates for institutional-portion deletion and matched random deletion. Rounded rates shown in the table can differ from the excess change rates computed on the raw data by 0.01 pp.

Model	Deleted portions	n	Portion deletion (%)	Random deletion (%)	Excess change rate (pp)	BH q
Llama	C/S/TS	1,893	6.66	6.87	−0.21 [−1.48, 1.06]	0.807
Qwen	C/S/TS	1,893	4.86	5.23	−0.37 [−1.43, 0.69]	0.675
Nemotron	C/S/TS	1,893	7.13	7.61	−0.48 [−1.85, 0.90]	0.675
Llama	U	959	10.74	9.28	1.46 [−0.73, 3.65]	0.670
Qwen	U	959	6.78	6.05	0.73 [−1.04, 2.50]	0.675
Nemotron	U	959	11.68	8.86	2.82 [0.42, 5.11]	0.130

With four classes retained, the excess change rates of C/S/TS portion deletion are 0.05 pp [−1.43, 1.53] for Llama, −0.11 pp [−1.27, 1.11] for Qwen, and −1.00 pp [−2.59, 0.58] for Nemotron, all with intervals including zero. For U portions, Nemotron shows a pre-correction difference in the three-class comparison, but q equals 0.130 after BH correction over the six comparisons. The absence of a detected difference is not converted into a claim that the effect is exactly zero or into an equivalence test.

Contrast with self-rationale deletion. For Llama, we extracted the first-rationale deletion of E.1 and its length-matched control on the same 1,893 documents as the C/S/TS portion deletion; the excess change rate is 9.56 pp [7.92, 11.25]. Qwen and Nemotron deleted their own first rationales separately over the 2,294-document intervention pool, with actual deletions and valid predictions on 2,152 and 1,936 documents respectively; their rationale and control change rates are 14.73% and 4.14% for Qwen, and 17.36% and 5.27% for Nemotron. The two extended-model columns of Table 5 come from the same data pool but are not re-paired on the identical documents holding

C/S/TS portions. Because the portion and self-rationale conditions differ in deletion scope and in the spans their controls avoid, the difference between the two excess change rates is not interpreted as a direct effect with deletion amount held fixed.