Evaluating LLM-Assisted Coding for Stable Narrative Representations in Web-Scale Fraud Analysis

Anonymous ACL submission

Abstract

Web-scale fraud analysis requires stable coding methods for transforming noisy web corpora into narrative representations capable of identifying emerging fraud patterns and broader social dynamics. However, traditional frequency-based methods, such as TF-IDF and LDA, often capture lexical salience rather than semantically meaningful fraud narratives, creating a representation mismatch between statistical features and narrative meaning. This study evaluates constrained LLM-assisted coding as a source-faithful narrative representation procedure for web-scale fraud analysis. We filter over 468 million raw records from Common Crawl and GNews to construct a fraud narrative corpus containing 160,248 fraud-related narrative paragraphs. We then extract source-faithful narrative units describing fraud schemes, methods, tools, roles, and behaviors. Compared with TF-IDF and LDA baselines, constrained LLM-assisted coding produces representations that are more closely aligned with narrative semantics and human judgments. Stability experiments further show that prompt and model variations affect local paragraph-level outputs, while the global narrative representation remains relatively stable under appropriate data filtering and constrained coding. These findings suggest that constrained LLM-assisted coding can serve as a feasible representation procedure for noisy web-scale corpora, providing a reproducible basis for computational narrative analysis and evolving fraud narratives.

1 Introduction

Effective crime governance and the maintenance of social stability depend on the timely collection and organization of data for understanding the conditions and evolving trends of criminal activities. With rapid digitalization, technological development has accelerated the evolution of criminal methods and reshaped the structure of criminal opportunities, giving rise to diverse cyber-enabled and emerging forms of crime (Wu et al., 2023; Onwuadiamu, 2025).

Under such dynamic conditions, the policy cycle from problem recognition to governance response faces substantial temporal pressure. Existing studies have shown that cybercrime-related legislation and governance systems often operate reactively, struggling to keep pace with rapidly changing criminal tactics (Khan et al., 2022; Chimchiuri, 2024). Consequently, both micro-level case observations and macro-level statistics face limitations in tracking the evolution of crime narratives. Web-based data sources have therefore been proposed as a way to compensate for the limits of traditional data in capturing dynamic social processes (Williams et al., 2016; Gyódi et al., 2022).

This issue is especially important for computational sociology and computational criminology, where large-scale data and artificial intelligence have been increasingly used to observe social phenomena, detect patterns, and support interpretation (Conte et al., 2012; Williams and Burnap, 2015; Smith et al., 2017; Edelmann et al., 2020). In this methodological context, Nelson introduced *Computational Grounded Theory (CGT)* (Nelson, 2017), which integrates natural language processing with grounded theory by using topic modeling to generate candidate themes from large-scale corpora before qualitative interpretation. CGT substantially expands the scale at which qualitative inquiry can operate (Tschisgale et al., 2023). However, scaling qualitative analysis also introduces a fundamental problem: the analytical pipeline must transform large, noisy, and heterogeneous textual data into stable and interpretable codes.

Prior work has noted that data-driven approaches may reduce contextual depth when they rely too heavily on statistical patterns (Baden et al., 2021). For example, frequency-based methods such as topic modeling and TF-IDF offer scalable and transparent means to identify salient lexical patterns,

but they often produce noisy or incoherent topics that fail to capture semantically meaningful units (Carlsen and Ralund, 2022; Nguyen et al., 2025). This occurs because statistical frequency does not inherently correspond to narrative importance (Gao et al., 2024). Related studies have emphasized the need to integrate qualitative and quantitative approaches more carefully (Macanovic, 2022; McCarthy and Maria, 2023). Such limitations become particularly consequential in crime narrative analysis, which requires representations that capture semantically meaningful units, including criminal methods, offender roles, deceptive tools, behavioral patterns, situational contexts, and institutional responses. These units describe how fraud is organized, adapted, and embedded in broader social and institutional contexts. As a result, traditional frequency-based methods create a representation-level mismatch between corpus-level statistical features and semantic-level crime narratives.

To address this mismatch, large language models (LLMs) have been introduced into grounded theory and narrative analysis as potential tools for contextually and semantically informed coding and interpretation (Sinha et al., 2024). By encoding contextual information and following task-specific instructions, LLMs offer a possible way to identify semantically meaningful narrative units rather than merely frequent lexical patterns. Yet web-based data are often fragmented, noisy, biased, duplicated, and may contain inappropriate or irrelevant content (Luccioni and Viviano, 2021). Under such conditions, LLM-assisted coding does not automatically guarantee stable and meaningful narrative representations, leaving important challenges in coding reliability and narrative interpretability unresolved (Olteanu et al., 2019; Daikeler et al., 2024).

These unresolved issues raise two methodological challenges for narrative coding in web-scale corpora. The first is whether a coding method can recover a stable conceptual structure from large-scale noisy corpora, rather than merely producing fragmented lexical or topic patterns. The second, specific to LLM-assisted coding, is whether the resulting codes remain stable across repeated runs, prompts, and model choices rather than reflecting coding artifacts. This study therefore asks: **How can stable crime narrative codes be constructed from large-scale noisy web corpora?**

We examine this question through fraud narrative analysis because fraud methods evolve rapidly across technological, financial, and social contexts, making their narratives difficult to capture through static taxonomies or purely frequency-based representations. To operationalize this question, we filter over 468 million raw records from Common Crawl (CC)[1] and GNews[2] to construct a fraud narrative corpus containing 160,248 fraud-related narrative paragraphs. We then build the coding datasets by extracting source-faithful narrative units describing fraud schemes, methods, tools, roles, and behaviors.

We formulate LLM-assisted coding as a constrained extraction task, requiring models to extract source-faithful narrative units related to fraud methods, tools, and behaviors. The extracted units are compared with TF-IDF and LDA baselines, evaluated against human judgments, and examined for stability across repeated runs, prompt variations, and model variations. This design allows us to assess whether LLM-assisted coding can produce meaningful local codes while preserving stable global narrative representations from large-scale noisy corpora.

Our contributions are summarized as follows:

- We frame web-scale crime narrative analysis as a representation problem involving a mismatch between statistical features and semantic narratives.
- We formulate LLM-assisted coding as a constrained extraction task and construct a large-scale fraud narrative corpus to validate.
- We develop an evaluation framework for comparing coding quality and stability across baselines, runs, prompts, and models.

2 Related Work

2.1 Narrative Analysis and Coding

Narrative analysis examines how events, actors, actions, and contexts are organized into meaningful accounts. Coding is central to this process because it transforms textual materials into analyzable concepts, categories, or themes while preserving their connection to the original empirical context. Although qualitative coding provides contextual sensitivity and interpretive depth, it becomes difficult to scale when corpora are large, heterogeneous, and continuously changing.

Computational approaches have therefore been introduced to support large-scale text analysis.

[1] https://commoncrawl.org/get-started
[2] https://gnews.io/

Methods such as topic modeling and clustering can generate candidate themes, group related documents, and help researchers explore large textual collections. Topic modeling has also been used in criminology and related text-based research (Kuang et al., 2017; K and Thilagam, 2019; Siddiqui et al., 2019; Saraiva et al., 2022; Tun and Birks, 2023). However, prior work emphasizes that computational text analysis still requires interpretation and validation to ensure that generated categories are meaningful for the research task (Grimmer and Stewart, 2013).

A key concern is that computationally generated patterns do not automatically correspond to narrative codes. Topic candidates may be sensitive to document length, dataset size, vocabulary sparsity, and other corpus-level conditions (Stevens et al., 2012), and statistically coherent topics are not always semantically interpretable to users (Chang et al., 2009). This study therefore treats coding as a representation problem and evaluates whether computational coding methods can produce source-faithful and semantically useful narrative units at web scale.

2.2 Web Corpora for Social Science

For narrative-oriented social science research, large-scale web corpora provide an important empirical resource because they capture how public discourse, media reporting, institutional communication, and other forms of socially produced text circulate across online environments. Compared with small manually collected datasets, web corpora allow researchers to observe narratives across wider temporal, geographic, and topical ranges. This is particularly valuable because narratives often emerge, diffuse, and change across distributed sources before they become stable taxonomies.

CC is one of the most representative open web corpora for this purpose. Through large-scale automated web crawling, it provides a low-cost and high-coverage source of textual data. However, CC should not be treated as a ready-to-use social science dataset. Baack notes that CC has several limitations for scientific research (Baack, 2024), including unstructured sampling, duplication, omission, missing or imprecise timestamps, restrictions caused by robots exclusion rules, and uneven language distribution. These limitations mean that CC is better understood as a collection of web snapshots than as a continuous and comprehensive representation of social dynamics. Web data are also noisy and may contain irrelevant, biased, or inappropriate content, making direct use problematic for scientific analysis.

Despite these limitations, CC can still serve as a source for narrative analysis when transformed through rigorous filtering and cleaning procedures. Prior work suggests that large web corpora can support strong semantic representation when appropriate filtering pipelines are applied (Weber et al., 2023; Kargaran et al., 2024; Su et al., 2025). Targeted or on-demand crawling can also supplement missing or underrepresented materials in open web corpora (Artetxe et al., 2022). In this sense, raw web data can provide partial but meaningful access to social reality when processed through transparent and reproducible procedures that extract relevant, source-faithful, and analyzable textual units.

To reduce noise and obtain more narratively meaningful materials, this study restricts the web corpus to online news content. Compared with general raw web pages, online news articles are more structured and undergo editorial filtering, making them more suitable for observing public representations of crime. News media may function both as a mirror of public thought and as a molder of public opinion (Peter, 2022). Although news reporting does not replace official statistics and may contain reporting biases, it remains an important source through which the public perceives crime risk and social problems (Jayaweera et al., 2015; Velásquez et al., 2020; Andersson, 2025).

For fraud analysis, this source is especially valuable because fraud methods evolve rapidly across technological, financial, and social contexts. Such evolution creates concept drift (Şenol et al., 2026), making static taxonomies and topic-based analysis difficult to maintain. The timely and high-frequency nature of news data makes them useful for tracking emerging criminal methods, public sentiment, and changing social issues (Iglesias et al., 2016; Krawczyk et al., 2021; Gokhberg et al., 2022). These properties motivate the use of both CC and GNews in this study, while also requiring a strict filtering pipeline before narrative coding.

2.3 LLMs for Qualitative Coding

LLMs as coding tools have become increasingly relevant to text analysis. Related developments in transformer-based topic modeling, contextual embeddings, and semantic representation suggest that contextualized models can capture meaning beyond surface frequency and word co-occurrence

(Grootendorst, 2022; Azher et al., 2024; Liu et al., 2025; Zhang et al., 2025). Previous studies also suggest that LLM-assisted coding can achieve similarity to human expert coding in social science and news-processing tasks (Than et al., 2025; Li et al., 2025; Ma et al., 2025).

The potential advantage of LLM-assisted coding lies in its ability to perform constrained semantic extraction. Instead of asking the model to freely summarize or interpret a document, researchers can define coding rules that require extracted units to remain faithful to the source text and relevant to a specific analytical purpose. If LLM outputs remain stable under such constraints, LLMs can support the transformation of large-scale corpora into candidate narrative units while preserving traceability to the original corpus. This use differs from generative interpretation because the model is not asked to replace human analysis, but rather to support analyzable narrative representations.

The central question, however, is whether such constrained outputs are sufficiently reliable for systematic analysis. Model outputs may vary across repeated runs, prompt formulations, and model families. In qualitative research, such variation cannot be ignored, because coding is meaningful only when it can support systematic comparison and interpretation.

Existing work has begun to examine the use of LLMs for qualitative and social science coding, suggesting that LLMs can reproduce several patterns of human language processing or show substantial agreement with human judgments in open-ended tasks (Zheng et al., 2023; Cai et al., 2024; Szymanski et al., 2025). Nevertheless, semantic quality alone does not automatically guarantee reproducibility. This study therefore evaluates LLM-assisted coding along two dimensions: alignment with human judgments and stability across runs, prompts, and model families.

3 Corpus Construction and Filtering

3.1 Data Sources

The experiment dataset was constructed by integrating collect-based and search-based acquisition strategies. CC serves as the collect-based source, providing broad web-scale coverage but also introducing substantial noise, duplication, and heterogeneous content that require extensive filtering. Because such broad crawling may still miss or underrepresent fraud-related news, we supplement it with GNews as a search-based source. GNews retrieves targeted fraud-news records through fraud-related queries, improving topical precision and temporal coverage. Combining these sources allows the dataset to balance web-scale breadth with targeted retrieval, thereby providing broader coverage of fraud-related news narratives from 2024 to 2025. Details of the data pipeline are provided in Appendix Table 5.

3.2 Document-Level Filtering

The first stage applies document-level filtering. For CC, we begin with language filtering and retain English content to ensure sufficient linguistic consistency. URL filtering is then used to restrict the corpus to domains containing the term "news." We further apply rule-based NLP filters based on keyword frequency, keyword density, and perplexity to remove low-relevance and low-quality documents. Finally, an LLM-based classifier is used to verify whether each candidate document contains fraud-related news content.

For GNews, the data are already retrieved through fraud-related search queries and are implicitly English. We therefore apply URL filtering, NLP-based filtering, and LLM-based relevance verification to ensure consistency with the CC pipeline.

3.3 Paragraph-Level Filtering

The second stage adopts a paragraph-centered filtering strategy. News webpages often contain navigation menus, advertisements, search engine optimization text, and other boilerplate content that can distort narrative analysis. Therefore, each document is segmented into paragraphs, and paragraphs are treated as the basic narrative records for subsequent coding.

Paragraph-level records are filtered using NLP-based rules that prioritize low perplexity and high keyword density. These rules are followed by LLM-based verification to confirm relevance to fraud-related content. This design reduces webpage-level noise and ensures that the analytical units used for coding are coherent, relevant, and narratively meaningful.

To support regional comparison, we further apply named entity recognition (NER) using the `spacy/en_core_web_trf` model to identify entities associated with twelve target regions. Documents are assigned to regions according to the frequency of matched entity occurrences.

Table 1: Corpus refinement pipeline for CC and GNews.

Stage	CC	GNews
Raw source	468.41M	29,689
Language	177.05M (37.8%)	–
URL	7.70M (1.64%)	14,756 (49.7%)
NLP	139,812 (0.03%)	4,583 (15.4%)
LLM	51,440 (0.001%)	2,727 (9.2%)

3.4 Filtering Results

As shown in Table 1, the filtering pipeline removes the vast majority of irrelevant records. From 468.41 million CC records, the pipeline retains 51,440 fraud-related news documents after LLM verification, corresponding to approximately 0.001% of the raw source. From 29,689 GNews records, 2,727 documents are retained after filtering. These results show that both broad web-scale sources and targeted news-search sources contain substantial noise, although in different forms and degrees. Large-scale fraud narrative analysis therefore requires appropriate filtering before coding, in order to preserve fraud-related narrative content while reducing irrelevant or weakly related records.

4 LLM-assisted Narrative Coding

After corpus construction and filtering, we formulate narrative coding as a constrained extraction task. The goal is to transform fraud-related narrative paragraphs into semantically meaningful and source-faithful narrative units. The coding method extracts a vocabulary of narrative units $\mathcal{V} = \{t_1, t_2, \ldots, t_d\}$ from the filtered paragraphs, where each t_i is a minimal semantic unit related to fraud methods, tools, behaviors, roles, infrastructures, or institutional actions. Each extracted unit must be grounded in the original paragraph text to preserve source faithfulness.

4.1 Constrained Narrative Unit Extraction

Traditional frequency-based methods often identify statistically salient terms, but these terms may not correspond to meaningful crime narratives. For example, general terms such as "fraud" or high-frequency function phrases may dominate the representation while offering limited analytical value. To reduce this mismatch, we use LLMs as constrained coding tools. The model is instructed to extract unigram, bigram, and trigram narrative units that are semantically related to fraud narratives.

This design treats LLM-assisted coding as a controlled semantic extraction process rather than unconstrained generation. The output is not a summary or interpretation, but a compact set of source-grounded narrative units that can support large-scale comparison and stability analysis.

Table 2: Corpus statistics across regions. Narrative units are source-faithful unigrams, bigrams, and trigrams extracted from filtered fraud-related paragraphs.

Metric	US	UK	IN	Others
Documents	32,956	7,009	6,839	7,363
Paragraphs	97,633	21,496	19,640	21,479
Unigrams	60,756	14,545	14,512	14,986
Bigrams	211,006	46,403	47,884	49,688
Trigrams	101,762	20,357	21,342	21,652
Narratives	373,524	81,305	83,738	86,326

4.2 Narrative Representation

The extracted narrative units define a shared semantic representation space. For each method m and condition $c \in \mathcal{C}$, such as source, year, or region, we construct a narrative representation vector $\mathcal{R}_c^{(m)} \in \mathbb{R}^d$, where the i-th dimension corresponds to a narrative unit $t_i \in \mathcal{V}$, and $\mathcal{R}_c^{(m)}[i]$ denotes the weight of t_i under condition c.

The global representation $\mathcal{R}^{(m)}$ is constructed from the full corpus and serves as the overall narrative baseline. Conditioned representations $\mathcal{R}_c^{(m)}$ are constructed from subsets of the corpus and are used to examine how fraud narratives vary across sources, years, and regions. All representations are computed over the full vocabulary $\mathcal{V}$, while Top-K narrative units are used only as an interpretable summary for observing patterns across conditions.

As shown in Table 2, the coding process extracts 624,893 narrative units from 160,248 narrative paragraphs across twelve target regions.

4.3 Narrative Stability Definition

We define narrative stability as the consistency of representation structures across analytical conditions. For each method m, extracted units are organized into a shared representation space. The Top-K units are selected from the global representation of method m using corpus-level TF-IDF weights and fixed as the common basis for all condition-level comparisons. Thus, when comparing two conditions c_1 and c_2, we retrieve the weights of the same Top-K units from each condition-specific representation rather than selecting separate Top-K units for each condition.

For two conditions c_1 and c_2, stability under method m is measured by the Pearson correlation between their vectors over the same Top-K units:

$$S_K^{(m)}(c_1, c_2) = \mathrm{corr}\left(\mathcal{R}_{c_1,K}^{(m)}, \mathcal{R}_{c_2,K}^{(m)}\right).$$

Here, $\mathcal{R}_{c,K}^{(m)}$ denotes the condition-specific TF-IDF weight vector restricted to the global Top-K units of method m. This fixed-basis design ensures that stability reflects changes in the relative importance of the same narrative units across conditions. The rationale is that a coding task should capture a shared conceptual narrative: although a concept such as fraud may vary in tactics or context, it should not depart entirely from the broader narrative structure of fraud. Stability is therefore assessed by whether condition-specific representations preserve this shared global narrative basis.

5 Experiment Results

5.1 Narrative Stability

To compare global representational robustness, we evaluate LLM-assisted coding using `Qwen3-30B-A3B-Instruct-2507` (Yang et al., 2025) against TF-IDF (Salton and Buckley, 1988) and LDA (Blei et al., 2003) under the proposed fixed-basis evaluation. Fig. 1 reports the resulting cross-condition stability. Source, year, and region-level comparisons show high corpus-level consistency, with median Pearson correlations above 0.75. LLM-assisted coding reaches stability comparable to the baselines, suggesting that semantically constrained extraction can preserve global representational robustness. In lower-data conditions, such as GNews subsets, it shows stronger stability, indicating lower sensitivity to data sparsity.

Fig. 2 further evaluates whether the Global Top-500 units extracted by LLM-assisted coding form a coherent representation rather than random artifacts. Across Top-K settings, all methods rapidly increase paragraph coverage, while LLM-assisted coding maintains the strongest global narrative similarity. This pattern suggests that LLM-assisted coding can provide broad corpus coverage while preserving a stable global narrative basis.

5.2 Semantic Relevance

Although TF-IDF and LDA also show high global stability, their stable features are not necessarily narratively meaningful. As shown in Fig. 3, the Top-30 features produced by LDA and TF-IDF include function words and fragmented phrases, whereas LLM-assisted coding produces units with clearer semantic content, including crime types, criminal methods, and legal or institutional terms.

This result highlights a key distinction between statistical stability and narrative validity. Frequency-based methods may produce stable lexical patterns, but such stability can reflect repeated linguistic forms rather than meaningful crime narratives. In this scenario, LLM-assisted coding provides a representation space that is more closely aligned with the semantic objects targeted by the coding task.

Table 3: Validation results after rater filtering. Lower Rank and higher Win indicate better performance.

Group	Metric	LLM	TF-IDF	LDA	Rand.
Police	Rank ↓	**1.585**	2.633	2.569	3.214
	Win ↑	**0.714**	0.088	0.113	0.085
Civil.	Rank ↓	**1.419**	2.664	2.631	3.286
	Win ↑	**0.764**	0.061	0.097	0.078
LLM	Rank ↓	**1.040**	2.620	2.860	3.480
	Win ↑	**0.960**	0.020	0.020	0.000

5.3 Human Validation

To evaluate whether the extracted narrative units align with human judgment, we conduct a ranking-based validation task. Participants compare candidate narrative sets generated by LLM-assisted coding, TF-IDF, LDA, and a Random baseline, and rank them according to their representativeness for a given paragraph. Participants include a professional group with police background and a civilian control group. A separate LLM-based evaluator is also included as an auxiliary comparison. Low-quality responses are filtered using a Random-first threshold of 0.15.

As shown in Table 3, after rater quality filtering, 32 of 38 human annotators were retained, including 17 police participants and 15 civilian participants (49 valid paragraphs; Police: $n = 833$; Civilian: $n = 735$). LLM-assisted coding achieves the best mean rank and highest win rate across all evaluator groups, outperforming TF-IDF, LDA, and Random. Friedman tests confirm significant differences among methods for all groups: Police, $\chi^2 = 82.48$, $p < 10^{-17}$; Civilian, $\chi^2 = 87.62$, $p < 10^{-18}$; and LLM, $\chi^2 = 97.08$, $p < 10^{-20}$. Moderate inter-rater agreement is observed for both human groups, with Kendall's $W = 0.57$ for Police and $W = 0.59$ for Civilian.

These results support the semantic validity of LLM-assisted coding.Combined with the stability

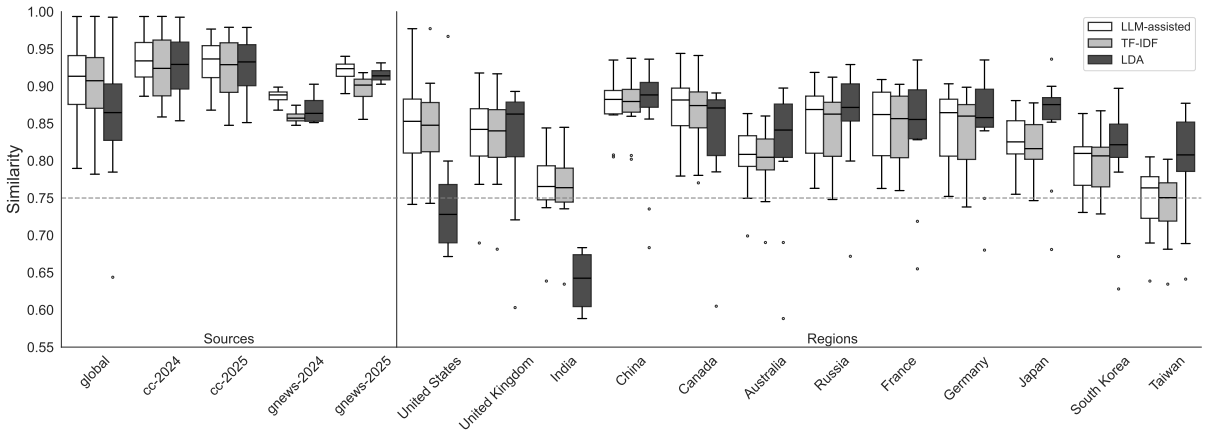

Figure 1: Stability comparison across coding methods using Top-K (K=500) global narrative units.

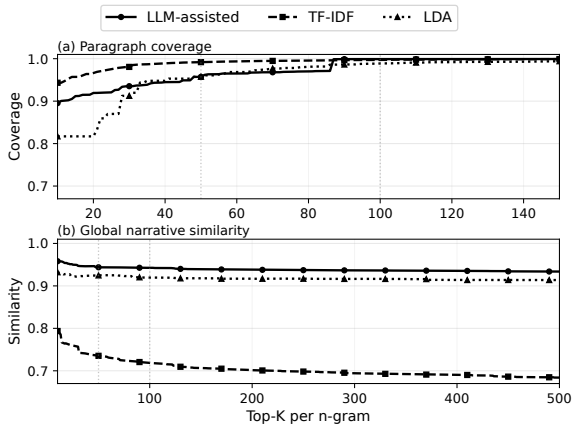

Figure 2: Paragraph-level coverage and global narrative similarity across Top-K selections.

analysis, the results suggest that LLM-assisted coding is globally stable and more closely aligned with human judgments of narrative relevance.

5.4 Prompt, Model, and Sampling Stability

Beyond comparing representation quality, we further examine whether the coding procedure is reproducible across repeated runs, prompt variations, and model choices. We therefore conduct three stability comparisons. The *same-condition* setting compares repeated runs using the same model and prompt. The *cross-prompt* setting compares different prompt formulations within the same model. The *cross-model* setting compares different LLMs under the same prompt. All comparisons are conducted on 160,248 fraud-related paragraphs (details see Appendix Table 9).

As shown in Table 4, stability is evaluated at both local and global levels. Paragraph similarity measures paragraph-level consistency using pairwise Jaccard similarity over extracted units. Pearson correlation measures global narrative stability by comparing the weight distributions of the same representation space across conditions. Jaccard similarity measures global set stability by evaluating the overlap of Top-K narrative units.

The results show a clear hierarchy of stability. Same-condition comparisons are the most stable, indicating that repeated coding under identical settings produces consistent local outputs and nearly identical global structures. Prompt variation lowers paragraph-level similarity, but the high global Pearson correlation suggests that the aggregated narrative distribution is largely preserved. Cross-model comparisons produce the largest local variation, showing that model choice affects paragraph-level coding more strongly than prompt variation, while still retaining a partially shared global structure (details see Appendix Fig. 5).

Fig. 4 further examines sampling stability. Same-condition stability remains consistently high across sample sizes. Under the cross-prompt setting, paragraph similarity is lower, while Pearson correlation remains high, suggesting that the overall narrative distribution is relatively stable despite prompt variation. Jaccard similarity is more sensitive to sample size, especially in smaller samples, indicating that Top-K membership is less stable than the overall weight distribution.

These results support a multi-level interpretation of LLM-assisted coding stability. Local outputs may vary across prompts and models, while narrative structures can remain relatively stable at the corpus level. LLM-assisted coding is therefore better understood as a representation process rather than a deterministic paragraph-labeling task.

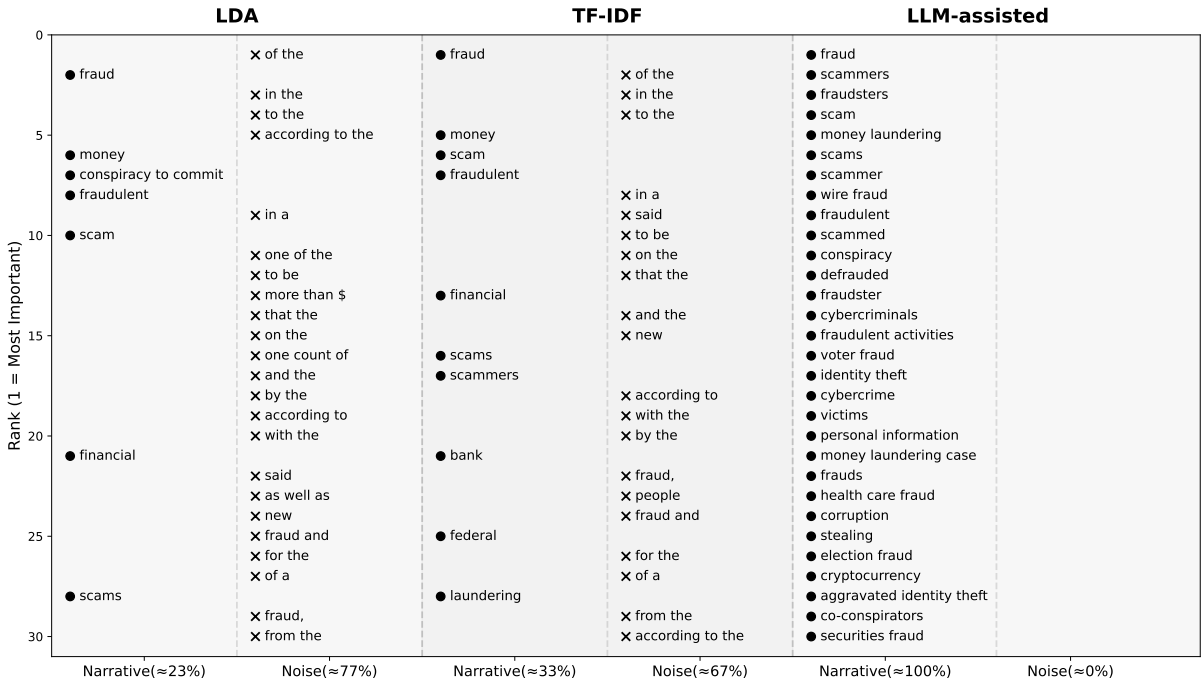

Figure 3: Top-30 features categorized as narrative-relevant or noise. LDA and TF-IDF show high noise rates, whereas LLM-assisted coding produces more semantically meaningful narrative units.

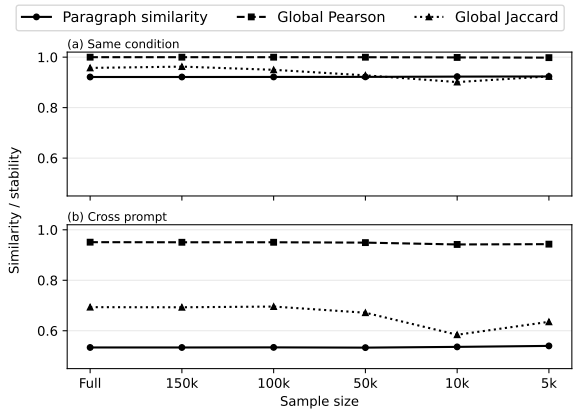

Figure 4: Sampling stability under same condition and cross prompt.

Table 4: Coding stability across repeated runs, prompt variants, and model variants. Paragraph similarity captures local coding consistency, Pearson captures global representational stability, and Jaccard captures overlap in the global Top-K narrative units ($K = 500$).

Type	Para. Sim.	Pearson	Jaccard
Same setting	0.925	0.999	0.969
Prompt variant	0.633	0.960	0.720
Model variant	0.465	0.917	0.581

6 Conclusion

This study examined whether constrained LLM-assisted coding can construct stable narrative representations from large-scale noisy web corpora. Using fraud narrative analysis as the empirical domain, we compared LLM-assisted coding with TF-IDF and LDA baselines and evaluated its semantic relevance, human alignment, and stability across analytical conditions.

The results provide three main findings. First, within the evaluated setting, LLM-assisted coding produces narrative units that are more closely aligned with the semantic objectives of narrative analysis than frequency-based baselines. Second, the extracted fraud narratives retain a shared global representation across sources and regional subsets, suggesting that the resulting units are not random coding artifacts. Third, although paragraph-level outputs may vary across prompts and models, the global narrative representation remains relatively stable under appropriate data filtering and constrained coding.

This study also contributes a reproducible pipeline for large-scale fraud narrative analysis. Overall, the findings support constrained LLM-assisted coding as a feasible representation procedure for transforming noisy web corpora into semantically meaningful and relatively stable narrative representations. By providing a quantifiable and reproducible basis for identifying narrative patterns, this approach can further support subsequent qualitative interpretation, comparative analysis, and domain-specific explanation.

Limitations

This study has several limitations. First, the comparison is limited to traditional frequency-based methods and constrained LLM-assisted coding; hybrid approaches combining frequency signals, contextual embeddings, or semantic clustering may also capture meaningful narrative structures. The extracted narrative units support large-scale comparison, but substantive interpretation still requires domain expertise. Third, because the corpus relies on web-derived materials, raw or reconstructable textual records are not publicly released due to possible third-party rights and redistribution restrictions. Reproducibility is instead supported through documentation of the pipeline, prompts, model settings, evaluation procedures, aggregate statistics, and computational environment. Finally, the model comparison is limited to the evaluated models and prompts, and the temporal separation of Common Crawl records is approximate. Future work should examine broader model families, hybrid representation methods, more precise temporal filtering, domain-specific coding criteria, and the interpretive process through which extracted narrative units are connected to substantive analytical meaning.

Acknowledgments

We thank the Common Crawl team for providing the large-scale web data used in this study, and the participants who contributed to the human validation experiment.

Ethical Considerations

The human validation experiment was conducted anonymously and focused only on information-processing preferences, not individual profiling. No personally identifiable information, biometric data, or sensitive demographic attributes were collected, and participation was voluntary and unpaid.

References

Dennis Andersson. 2025. News media use and crime perceptions: The dual role of ideology. *Mass Communication and Society*, 28:511–529.

Mikel Artetxe, Itziar Aldabe, Rodrigo Agerri, Olatz Perez de Viñaspre, and Aitor Soroa. 2022. Does corpus quality really matter for low-resource languages? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, page 7383–7390. Association for Computational Linguistics.

Ibrahim Al Azher, Venkata Devesh Reddy Seethi, Akhil Pandey Akella, and Hamed Alhoori. 2024. Limtopic: Llm-based topic modeling and text summarization for analyzing scientific articles limitations. In *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*, pages 1–12. Association for Computational Linguistics.

Stefan Baack. 2024. A critical analysis of the largest source for generative ai training data: Common crawl. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*.

Christian Baden, Christian Pipal, Martijn Schoonvelde, and Mariken A. C. G van der Velden. 2021. Three gaps in computational text analysis methods for social sciences: A research agenda. *Communication Methods and Measures*, 16:1–18.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Zhenguang Cai, Xufeng Duan, David Haslett, Shuqi Wang, and Martin Pickering. 2024. Do large language models resemble humans in language use? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 37–56. Association for Computational Linguistics.

Hjalmar Bang Carlsen and Snorre Ralund. 2022. Computational grounded theory revisited: From computer-led to computer-assisted text analysis. *Big Data Society*, 9:205395172210801.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-graber, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, pages 288 – 296. Association for Computational Linguistics.

Lika Chimchiuri. 2024. The evolution of cybercrime legislation. *Scientific works of National Aviation University. Series: Law Journal "Air and Space Law"*, 2:221–227.

R. Conte, N. Gilbert, G. Bonelli, C. Cioffi-Revilla, G. Deffuant, J. Kertesz, V. Loreto, S. Moat, J. P. Nadal, A. Sanchez, A. Nowak, A. Flache, M. San Miguel, and D. Helbing. 2012. Manifesto of computational social science. *The European Physical Journal Special Topics*, 214:325–346.

Jessica Daikeler, Leon Fröhling, Indira Sen, Lukas Birkenmaier, Tobias Gummer, Jan Schwalbach, Henning Silber, Bernd Weiß, Katrin Weller, and Clemens Lechner. 2024. Assessing data quality in the age of digital social research: A systematic review. *Social Science Computer Review*, 43.

Achim Edelmann, Tom Wolff, Danielle Montagne, and Christopher A. Bail. 2020. Computational social science and sociology. *Annual Review of Sociology*, 46:61–81.

Sally Gao, Milda Norkute, and Abhinav Agrawal. 2024. Evaluating interactive topic models in applied settings. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–8.

Leonid Gokhberg, Dirk Meissner, and Ilya Kuzminov. 2022. What semantic analysis can tell us about long term trends in the global sti policy agenda. *The Journal of Technology Transfer*, 48:2249–2277.

Justin Grimmer and Brandon M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21:267–297.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure.

Kristóf Gyódi, Łukasz Nawaro, Michał Paliński, and Maciej Wilamowski. 2022. Informing policy with text mining: Technological change and social challenges. *Quality Quantity*, 57.

José Antonio Iglesias, Alexandra Tiemblo, Agapito Ledezma, and Araceli Sanchis. 2016. Web news mining in an evolving framework. *Information Fusion*, 28:90–98.

Isuru Jayaweera, Chamath Sajeewa, Sampath Liyanage, Tharindu Wijewardane, Indika Perera, and Adeesha Wijayasiri. 2015. Crime analytics: Analysis of crimes through newspaper articles. In *2015 Moratuwa Engineering Research Conference*. IEEE.

Srinivasa K and P. Santhi Thilagam. 2019. Crime base: Towards building a knowledge base for crime entities and their relationships from online news papers. *Information Processing Management*, 56:102059.

Amir Kargaran, Hinrich Schütze, and François Yvon. 2024. Glotcc: An open broad-coverage commoncrawl corpus and pipeline for minority languages. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 16983–17005.

Shereen Khan, Tajneen Saleh, Magiswary Dorasamy, Nasreen Khan, Olivia Tan Swee Leng, and Rossanne Gale Vergara. 2022. A systematic literature review on cybercrime legislation. *F1000Research*, 11:971.

Konrad Krawczyk, Tadeusz Chelkowski, Daniel J. Laydon, Swapnil Mishra, Denise Xifara, Benjamin Gibert, Seth Flaxman, Thomas Mellan, Veit Schwämmle, Richard Röttger, Johannes T. Hadsund, and Samir Bhatt. 2021. Quantifying online news media coverage of the covid-19 pandemic: Text mining study and resource. *Journal of Medical Internet Research*, 23:e28253.

Da Kuang, P. Jeffrey Brantingham, and Andrea L. Bertozzi. 2017. Crime topic modeling. *Crime Science*, 6.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791. Association for Computational Linguistics.

Jianghan Liu, Ziyu Shang, Wenjun Ke, Peng Wang, Zhizhao Luo, Jiajun Liu, Guozheng Li, and Yining Li. 2025. Llm-guided semantic-aware clustering for topic modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 18420–18435. Association for Computational Linguistics.

Alexandra Luccioni and Joseph Viviano. 2021. What's in the box? an analysis of undesirable content in the common crawl corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, page 182–189. Association for Computational Linguistics.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. 2025. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 8679–8696. Association for Computational Linguistics.

Ana Macanovic. 2022. Text mining for social science – the state and the future of computational text analysis in sociology. *Social Science Research*, 108:102784.

Arya D McCarthy and Giovanna Maria. 2023. Theory-grounded computational text analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, volume 2. Association for Computational Linguistics.

Laura K. Nelson. 2017. Computational grounded theory. *Sociological Methods Research*, 49:004912411772970.

Quang Duc Nguyen, Tung Nguyen, Duc Anh Nguyen, Linh Ngo Van, Sang Dinh, and Thien Huu Nguyen. 2025. Glocom: A short text neural topic model via global clustering context. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 1109–1124. Association for Computational Linguistics.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2.

Gift Onwuadiamu. 2025. Cybercrime in criminology; a systematic review of criminological theories, methods, and concepts. *Journal of Economic Criminology*, 8:100136.

Christina Peter. 2022. Media coverage as mirror or molder? an inference-based framework. *Media and Communication*, 10:183–195.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing Management*, 24:513–523.

Miguel Saraiva, Irina Matijošaitienė, Saloni Mishra, and Ana Amante. 2022. Crime prediction and monitoring in porto, portugal, using machine learning, spatial and text analytics. *ISPRS International Journal of Geo-Information*, 11:400.

Tamanna Siddiqui, Abdullah Yahya Abdullah Amer, and Najeeb Ahmad Khan. 2019. Criminal activity detection in social network by text mining: Comprehensive analysis. In *2019 4th International Conference on Information Systems and Computer Networks*. IEEE.

Ravi Sinha, Idris Solola, Ha Nguyen, Hillary Swanson, and LuEttaMae Lawrence. 2024. The role of generative ai in qualitative research: Gpt-4's contributions to a grounded theory analysis. In *Proceedings of the 2024 Symposium on Learning, Design and Technology*, pages 17–25.

Gavin J. D. Smith, Lyria Bennett Moses, and Janet Chan. 2017. The challenges of doing criminology in the big data era: Towards a digital and data-driven approach. *The British Journal of Criminology*, 57:259–274.

Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, page 952–961. Association for Computational Linguistics.

Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2025. Nemotron-cc: Transforming common crawl into a refined long-horizon pretraining dataset. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 2459–2475. Association for Computational Linguistics.

Annalisa Szymanski, Noah Ziems, Heather A Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A Metoyer. 2025. Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 952–966. Association for Computational Linguistics.

Nga Than, Leanne Fan, Tina Law, Laura K. Nelson, and Leslie McCall. 2025. Updating "the future of coding": Qualitative coding with generative large language models. *Sociological Methods Research*, 54:849–888.

Paul Tschisgale, Peter Wulff, and Marcus Kubsch. 2023. Integrating artificial intelligence-based methods into qualitative research in physics education research: A case for computational grounded theory. *Physical Review Physics Education Research*, 19.

Zeya Lwin Tun and Daniel Birks. 2023. Supporting crime script analyses of scams with natural language processing. *Crime Science*, 12.

Daniel Velásquez, Santiago Medina, Gustavo Yamada, Pablo Lavado, Miguel Nunez-del Prado, Hugo Alatrista-Salas, and Juandiego Morzán. 2020. I read the news today, oh boy: The effect of crime news coverage on crime perception. *World Development*, 136:105111.

Maurice Weber, Carlo Siebenschuh, Rory Butler, Anton Alexandrov, Valdemar Thanner, Georgios Tsolakis, Haris Jabbar, Ian Foster, Bo Li, Rick Stevens, and Ce Zhang. 2023. Wordscape: A pipeline to extract multilingual, visually rich documents with layout annotations from web crawl data. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 26048–26068.

Matthew L. Williams and Pete Burnap. 2015. Cyberhate on social media in the aftermath of woolwich: A case study in computational criminology and big data. *British Journal of Criminology*, 56:211–238.

Matthew L. Williams, Pete Burnap, and Luke Sloan. 2016. Crime sensing with big data: The affordances and limitations of using open source communications to estimate crime patterns. *British Journal of Criminology*, 57:azw031.

Ling Wu, Qiong Peng, and Michael Lembke. 2023. Research trends in cybercrime and cybersecurity: A review based on web of science core collection database. *International Journal of Cybersecurity Intelligence and Cybercrime*, 6:5–28.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report.

Longyin Zhang, Bowei Zou, and AiTi Aw. 2025. Enhancing event-centric news cluster summarization via data sharpening and localization insights. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 16412–16426. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 46595 – 46623. Association for Computational Linguistics.

Ali Şenol, Garima Agrawal, and Huan Liu. 2026. Domain knowledge-enhanced llms for fraud and concept drift detection. *Electronics*, 15:534.

A Appendix

A.1 Computational Hardware

Corpus construction and filtering were conducted on an NVIDIA DGX Spark with 128 GB of memory. The stability analysis was conducted on a server equipped with 8 NVIDIA H200 GPUs.

Table 5: Multi-stage filtering and transformation pipeline for fraud corpus construction and narrative extraction. Each stage progressively enforces semantic relevance and structural consistency.

Phase	Stage	Process and Criteria
Raw	Input	Raw documents from Common Crawl (WET) and GNews sources.
Filtering	**Heuristic Filtering**	For each document d, retain English content, apply a URL whitelist (news domains), and compute $\mathrm{TF}(d)$ and density $\rho(d)$. Retain d if $\mathrm{TF}(d) \geq 5$, $\rho(d) \geq 0.003$, and the high-density paragraph ratio is ≥ 0.2, where $\rho(p) \geq 0.01$. *Purpose: enforce coarse semantic relevance to fraud discourse.*
	Perplexity Filtering	For each retained document d, compute perplexity using GPT-2 (max 512 tokens) and retain d if $\mathrm{PPL}(d) < 200$. *Purpose: remove noisy or non-linguistic content and ensure fluency.*
	LLM Verification (Prompt see Table 6)	For each remaining document d, apply LLM with verification prompt (`Qwen/Qwen3-30B-A3B-Instruct`, temperature=0.0, max tokens=200) and retain d if `is_fraud_news = true`. *Purpose: enforce fine-grained semantic validity.*
	Deduplication	For each document d, compute URL hash h_u and content hash h_c, and retain d if $h_u \notin S_u$ and $h_c \notin S_c$. *Purpose: remove duplicate or near-duplicate documents.*
	Delta Construction	Construct Δ_{2024}, Δ_{2025}, and $\mathcal{O}$ from U_{2024}, U_{2025}. *Purpose: isolate temporal variation and stable narratives.*
	NER & Country Attribution	For each document d, apply spaCy NER (`spacy/en_core_web_trf`) to extract geographic candidates from {ORG, GPE, LOC}, map them via an alias dictionary to canonical regions, and assign $\mathrm{region}(d) = \arg\max_c \mathrm{count}(c)$. *Purpose: enrich documents with structured semantic attributes for downstream analysis.*
	Document → Paragraph (Prompt see Table 7)	Split document d into paragraphs p_i using newline boundaries, retain paragraphs with $30 \leq \lvert p \rvert \leq 200$ tokens, compute fraud density $\rho(p)$, and retain p if $\rho(p) \geq 0.01$. Rank paragraphs by $\rho(p)$, keep top-K per document ($K = 6$), apply a GPT-2 perplexity filter retaining $\mathrm{PPL}(p) < 150$, and apply an LLM classifier (`Qwen/Qwen3-30B-A3B-Instruct`, temperature=0.0, max tokens=100) to retain `is_fraud_news = true`. *Purpose: select high-density, fluent, and semantically valid paragraphs for downstream analysis.*
Coding	**Paragraph → Narrative Units (Prompt see Table 8)**	For each paragraph p, use an LLM (`Qwen/Qwen3-30B-A3B-Instruct`, temperature=0.0, max tokens=200) to extract candidate n-grams ($n = 1, 2, 3$), and retain candidates that (1) appear exactly in the original text, (2) contain at most 3 tokens, and (3) are fraud-relevant; remove duplicates. *Purpose: construct compact, semantically grounded narrative units.*

Table 6: LLM prompt for fraud news verification.

System Prompt
You are a classification model for fraud-related news. Your task is to determine whether the given text is a news article that reports, describes, or investigates a concrete fraud incident. Output MUST be valid JSON in the following format: `{ "is_fraud_news": true/false, "reason": "short explanation" }` Guidelines: Return true ONLY if the text describes real fraud cases, schemes, investigations, or victims. Return false if the text is about fraud prevention, product announcements, corporate services, or general discussions. The reason must be concise, factual, and no longer than 20 words. Do NOT include opinions, speculation, or policy commentary.

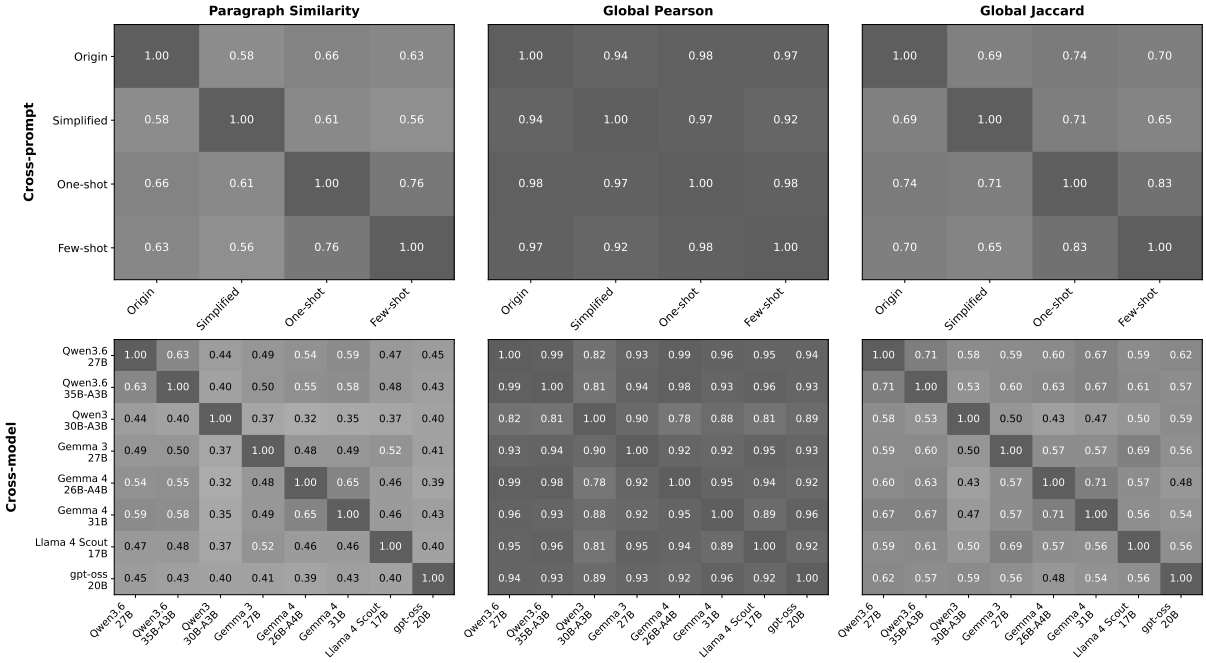

Figure 5: Aggregated cross-prompt and cross-model stability heatmaps. The upper row summarizes prompt-level variation, while the lower row summarizes model-level variation, using global Top-K narrative units ($K = 500$).

Table 7: Paragraph-Level LLM prompt for fraud event filtering.

System Prompt
You are a classification model for fraud-related news. Your task is to determine whether the given text is a meaningful news paragraph that directly participates in describing a concrete fraud incident, investigation, victimization, or operational scheme. Output MUST be valid JSON in the following format: `{ "is_fraud_news": true/false, "reason": "short explanation" }` Return TRUE ONLY if the paragraph: - Describes a real fraud case, scheme, investigation, or victim - Mentions concrete actors, actions, methods, or consequences - Contributes to the narrative of a specific fraud event Return FALSE if the paragraph is: - A product, company, or service announcement - A navigation item, site menu, or category label - A newsletter signup, subscription prompt, or contact info - A section header or title fragment - A very short teaser or summary lacking event detail - A clickbait lead-in or promotional paragraph SEO-generated preview text - A social media sharing prompt or engagement request - Any paragraph that does not contain a concrete fraud action Important: - Presence of fraud-related keywords alone is NOT sufficient. - The paragraph must describe or participate in an actual fraud event. - Ignore generic, templated, or structural webpage content. The reason must be concise (<=20 words). - Do NOT include opinions or speculation.

Table 8: LLM prompt for extracting fraud-related candidates

System Prompt

You extract fraud-related candidate phrases from a given paragraph.

TASK DEFINITION:
Your goal is to identify short phrases that are semantically relevant to FRAUD DISCOURSE STRUCTURE. This includes fraud acts, roles, schemes, operational methods, victims, or infrastructure that enable or support fraud operations.

INCLUDE phrases that describe:
- Fraud or scam acts (e.g. fraud, scam, phishing)
- Fraud schemes or case types (e.g. medicare fraud, phone scam)
- Operational methods or tools (e.g. stolen identities, fake accounts)
- Roles or victims involved in fraud ecosystems
- Infrastructure or enabling elements of fraud operations

EXCLUDE phrases that are:
- Monetary amounts, loss values, or financial damage descriptions
- Generic legal or administrative terms (e.g. conspiracy, indictment)
- Names of institutions, agencies, media outlets, or websites
- Procedural or abstract legal concepts without fraud specificity
- Paragraphs that only describe impact without method or actor

STRICT CONSTRAINTS:
- Every returned phrase MUST appear EXACTLY in the paragraph as a contiguous substring.
- Do NOT paraphrase, infer, normalize, or invent new terms.
- Each phrase must contain AT MOST 3 words.
- Prefer operational or role-related phrases over outcome descriptions.
- Monetary loss or damage information must NOT be extracted.
- If no valid fraud-related phrase exists, return an empty list.

OUTPUT FORMAT (JSON ONLY):
```
{ "candidates": ["exact phrase", "another phrase"] }
```

IMPORTANT:
- Do NOT include explanations or comments.
- Do NOT include duplicate phrases.

Table 9: Prompt, model, and repeated-run settings used for coding stability comparisons.

Setting	Variants	Purpose
Prompt variants	origin (Prompt see Table 8) simplified (Prompt see Table 10) one-shot (Prompt see Table 11) few-shot (Prompt see Table 12)	Tests whether coding remains stable when the task is expressed through different same-task prompt formulations.
Model variants	Qwen/Qwen3-30B-A3B-Instruct-2507 Qwen/Qwen3.6-27B Qwen/Qwen3.6-35B-A3B google/gemma-3-27b-it google/gemma-4-31B-it google/gemma-4-26B-A4B-it meta-llama/Llama-4-Scout-17B-16E-Instruct openai/gpt-oss-20b	Tests whether coding remains stable across different instruction-tuned LLMs.
Repeated runs	3 runs per prompt	Tests run-level stability under the same model–prompt condition.
Comparison types	Same condition: 96 pairs Prompt variant: 144 pairs Model variant: 336 pairs	Defines the pairwise comparisons used in the stability analysis.

Table 10: Simplified LLM prompt for extracting fraud-related candidates

System Prompt

Extract short phrases that are semantically relevant to FRAUD DISCOURSE STRUCTURE from the paragraph.

Include phrases describing fraud acts, schemes, roles, victims, operational methods, tools, or infrastructure/enabling elements of fraud operations.

Requirements:
- Each phrase must appear exactly in the paragraph as a contiguous substring.
- Each phrase must contain at most 3 words.
- Do not paraphrase, infer, normalize, or invent new terms.
- Prefer operational or role-related phrases over outcome descriptions.
- Do not include monetary values, loss amounts, financial damage descriptions, generic legal or administrative terms, institution/agency/media/website names, or procedural/abstract legal concepts without fraud specificity.
- Do not extract from paragraphs that only describe impact without method or actor.
- Do not include duplicate phrases.
- Return an empty list if no valid fraud-related phrase exists.

Return JSON only:
`{ "candidates": ["exact phrase", "another phrase"] }`

Table 11: One-shot LLM prompt for extracting fraud-related candidates

System Prompt

You extract short phrases that are semantically relevant to FRAUD DISCOURSE STRUCTURE from a paragraph.

Each phrase must:
- Appear exactly in the paragraph as a contiguous substring
- Contain at most 3 words
- Represent fraud acts, schemes, roles, victims, operational methods, tools, or infrastructure/enabling elements

Example:
Paragraph:
`"The suspect used fake accounts and phishing emails to conduct wire fraud against elderly victims."`

Output:
`{ "candidates": ["fake accounts", "phishing emails", "wire fraud", "elderly victims"] }`

Rules:
- Do NOT paraphrase, infer, normalize, or invent new terms.
- Do NOT include phrases longer than 3 words.
- Do NOT include monetary values, loss amounts, financial damage descriptions, generic legal or administrative terms, institution/agency/media/website names, or procedural/abstract legal concepts without fraud specificity.
- Prefer operational or role-related phrases over outcome descriptions.
- Do NOT include duplicate phrases.
- Return an empty list if no valid fraud-related phrase exists.

Return JSON only:
`{ "candidates": ["exact phrase", "another phrase"] }`

Table 12: Few-shot LLM prompt for extracting fraud-related candidates

System Prompt

You extract short phrases that are semantically relevant to FRAUD DISCOURSE STRUCTURE from a paragraph.

Each phrase must:
- Appear exactly in the paragraph as a contiguous substring
- Contain at most 3 words
- Represent fraud acts, schemes, roles, victims, operational methods, tools, or infrastructure/enabling elements

Example 1:
Paragraph:
`"The suspect used fake accounts and phishing emails to conduct wire fraud against elderly victims."`

Output:
`{ "candidates": ["fake accounts", "phishing emails", "wire fraud", "elderly victims"] }`

Example 2:
Paragraph:
`"Investigators said the group submitted false claims using stolen identities and recruited money mules to move the proceeds."`

Output:
`{ "candidates": ["false claims", "stolen identities", "money mules"] }`

Example 3:
Paragraph:
`"Victims were contacted through romance scams and directed to transfer funds through cryptocurrency wallets controlled by fraudsters."`

Output:
`{ "candidates": ["Victims", "romance scams", "cryptocurrency wallets", "fraudsters"] }`

Rules:
- Do NOT paraphrase, infer, normalize, or invent new terms.
- Do NOT include phrases longer than 3 words.
- Do NOT include monetary values, loss amounts, financial damage descriptions, generic legal or administrative terms, institution/agency/media/website names, or procedural/abstract legal concepts without fraud specificity.
- Prefer operational or role-related phrases over outcome descriptions.
- Do NOT include duplicate phrases.
- Return an empty list if no valid fraud-related phrase exists.

Return JSON only:
`{ "candidates": ["exact phrase", "another phrase"] }`

Human Validation for Fraud Related Narratives

Task Overview
In this task, you will evaluate several token sets extracted from a paragraph about fraud-related news.
Your goal is to determine which token set best captures the core concepts (Narrative Units) of the paragraph while avoiding unrelated or purely tokens.
You should rank all candidate sets from best to worst.

==========
What is a Narrative Unit?
A narrative unit is a n-gram word that represents a key concept necessary to understand the main event or mechanism described in the paragraph.
Examples of narrative units in fraud-related text:

fraud
scam
phishing
money laundering
victims
crypto
scheme
investigation

These tokens capture the core semantic elements of the event.

Figure 6: Instructions provided to participants in the human validation task. We used Google Forms to collect participants' rankings of candidate narrative units generated by different methods according to their representativeness for the given fraud-related paragraph.